Universidad de Málaga

Escuela Técnica Superior de Ingeniería de Telecomunicación

Programa de Doctorado en Ingeniería de Telecomunicación

TESIS DOCTORAL

Application of Harris Hawks Optimization (HHO) and Genetic Algorithms to Biodata Systems

Author:

Haedar Alsafi

Director:

Jorge Munilla

10 de octubre de 2024

UNIVERSIDAD DE MÁLAGA

AUTOR: Haedar Emad Sharef Al-Safi

https://orcid.org/0000-0001-9146-7762

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR

D./Dña HAEDAR EMAD SHAREF AL-SAFI
Estudiante del programa de doctorado PROGRAMA DE DOCTORADO EN INGENIERÍA DE TELECOMUNICACIÓN de la Universidad de Málaga, autor/a de la tesis, presentada para la obtención del título de doctor por la Universidad de Málaga, titulada: PROGRAMA DE DOCTORADO EN INGENIERÍA DE TELECOMUNICACIÓN

Realizada bajo la tutorización de JORGE MUNILLA y dirección de JORGE MUNILLA

DECLARO QUE:

La tesis presentada es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, conforme al ordenamiento jurídico vigente (Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), modificado por la Ley 2/2019, de 1 de marzo.

Igualmente asumo, ante a la Universidad de Málaga y ante cualquier otra instancia, la responsabilidad que pudiera derivarse en caso de plagio de contenidos en la tesis presentada, conforme al ordenamiento jurídico vigente.

En Málaga, a 04 de OCTOBER de 2024

Fdo.: Doctorando/a HAE DAR EMAD SHAREF AL-SAFI	Fdo.: Tutor/a Jorge Munilla Fajardo
Fdo.: Director/es de tesis Jorge Munilla Fajardo	

Edificio Pabellón de Gobierno. Campus El Ejido. 29071
Tel.: 952 13 10 28 / 952 13 14 61 / 952 13 71 10
E-mail: doctorado@uma.es

INFORME SOBRE LA REALIZACIÓN DE TRABAJOS BAJO MI TUTORIZACIÓN/SUPERVISIÓN Y NO UTILIZACIÓN DE ESTOS PARA LA DEFENSA DE OTRAS TESIS DOCTORALES

D. Jorge Munilla Fajardo

Profesor titular del Departamento de Ingeniería de Comunicaciones de la Universidad de Málaga, tutor del estudiante del programa de doctorado Haedar Emad Sharef Al-Safi, y director de la tesis, presentada por este para la obtención del título de doctor por la Universidad de Málaga, titulada:

Application of Harris Hawks Optimization (HHO) and Genetic Algorithms to Biodata Systems

DECLARO QUE:

Las siguientes publicaciones en conferencias internacionales y revistas que avalan la tesis:

Conferencias internacionales:

- H. Al-Safi, J. Munilla, and J. Rahebi, "Harris Hawks Optimization (HHO) Algorithm based on Artificial Neural Network for Heart Disease Diagnosis," in *2021 IEEE International Conference on Mobile Networks and Wireless Communications (ICMNWC)*, 2021, pp. 1–5.
- H. Alsafi, H. Alsalihi, and J. Munilla, "Use Harris Hawks Optimization (HHO) Algorithm based on Artificial Neural Network for liver disease diagnosis ," *Proc. Int. Conf. Intell. Syst. New Appl.*, vol. 2, no. SE-Proceedings Paper, pp. 1–10, Apr. 2024.

Revistas indexadas en el JCR:

- H. Al-Safi, J. Munilla, and J. Rahebi, "Patient privacy in smart cities by blockchain technology and feature selection with Harris Hawks Optimization (HHO) algorithm and machine learning," *Multimed. Tools Appl.*, pp. 1–25, 2022.
- H. Alsafi, J. Munilla, and J. Rahebi, "An Approach for Cardiac Coronary Detection of Heart Signal Based on Harris Hawks Optimization and Multichannel Deep Convolutional Learning," Comput. Intell. Neurosci., vol. 2022, 2022.
- J. Munilla, H. E. S. Al-Safi, A. Ortiz, and J. L. Luque, "Hybrid Genetic Algorithm for Clustering IC Topographies of EEGs," *Brain Topogr.*, vol. 36, no. 3, pp. 338–349, 2023.

Han sido realizadas, bajo mi dirección y tutorización, por D. Haedar Emad Sharef Al-Safi en colaboración conmigo y otros investigadores, y que estas no han sido, ni serán, utilizadas para avalar ningún otro trabajo doctoral.

En Málaga, a 3 de octubre de 2024

Fdo: Jorge Munilla Fajardo
Tutor y Director de la Tesis

Edificio Pabellón de Gobierno. Campus El Ejido. 29071
Tel.: 952 13 10 28 / 952 13 14 61 / 952 13 71 10
E-mail: doctorado@uma.es

Escuela de ETSI de Telecomunicación

Programa de Doctorado en Ingeniería de Telecomunicación

TESIS DOCTORAL

Application of Harris Hawks Optimization (HHO) and Genetic Algorithms to Biodata Systems

Aplicación de la Optimización de Harris Hawks (HHO) y Algoritmos Genéticos a Sistemas de Biodatos

Author:

Haedar Alsafi

Director:

Jorge Munilla

Ingeniería de Telecomunicación

UNIVERSIDAD DE MALAGA

10 de octubre de 2024

Agradecimientos

I express my deep appreciation to Professor Jorge Munilla, who was my advisor during my PhD studies as well as my research, who unceasingly made me feel motivated, encouraged, and directed. His comments and recommendations assisted in the course of the entire undertaking of this thesis project. I had the great fortune of having an advisor and mentor who was able to support and motivate me at each phase of my PhD studies.

ABSTRACT:

Biological data is inclusive of a plethora of information such as genomic data, biological signals, and clinical documents and records, including data acquisition technologies. With the progression of technology, biological informatics is regarded as one of the important means of growth in medicine and healthcare, agriculture, Neuroscience, and the environmental sciences. However, its usefulness is exceedingly challenged by the nature of its data, which is very intricate, noisy, and unreliable. Although useful, traditional data mining approaches are not able to get rid of such noise and deal with the volume and the level of intricacy, which calls for the use of more advanced computational strategies.

Biological AI data, machine learning, and deep learning, possess the qualities to handle Big data and specifically set queries aimed at specific big data. They can form relationships, patterns, and identify structures in large sets of data where using statistics alone would be impossible. In bioinformatics, it is critical to select a subset of features that reliably differentiates sets of highly complex biological data. There is no need for humans to do so since AI can find important features themselves and decrease dimensions thereby enhancing efficiency. The core topic of this work, explained more below, was to utilize Harris hawk optimization, which is an optimization algorithm to automatically select the most relevant features.

Artificial neural networks (ANNs) are computer programs that are designed to function in a manner similar to the neural networks present in the human brain. These consist of a series of layers of nodes (neurons) with each neuron receiving some input signals, performing a weighted sum, and transmitting it through a nonlinear transformation called an activation function. This organization of the network enables ANNs to learn intricate, nonlinear dependencies within their input, and thus, they have been successfully used in diagnosis of disease, prediction of protein structure, analysis of gene expression and many others.

One of the vital problems of modeling biological data systems based on ANNs is high dimensionality and intricacy of the data. ANNs may also overfit the training data or not generalize well if the data is not optimized correctly.

The Harris Hawk Optimization (HHO) algorithm was developed based on the cooperative behavior of the falcons to hunt while performing the two key tasks of seeking and capturing in unpredictable motions seeking food. All these activities can well be modeled and executed in optimization processes, in which the former blocks the algorithm from searching too broadly across the problem and the latter makes it possible to achieve a locally optimal solution. In more detail, the HHO algorithm incorporates different processes, such as initialization, exploration, exploitation, cooperative hunting, and updating of the position. In this stage of the model, the second task of capturing prey is not performed but instead falcons seek unoccupied areas in the search space. Then comes the exploitation phase in which the validated or selected improvement strategies are implemented to the set of promising areas as the falcons move to those areas. The cooperative hunting phase is essential especially for narrowing down the possible solution space, as it enables falcons to communicate amongst themselves and merges their intelligence to a single best solution. This element also avoids the scenario where the algorithm provides solutions that are not optimum and with regard to this tendency, HHO is considered a better algorithm for feature selection in high dimensionality and complex data set which are typical in biological data system.

An evolutionary algorithm called the Harris hawk optimization in conjunction with neural networks was advanced in this thesis and seeks for the specifics of causes of congestive heart failure or liver disease among various things. Neural network incorporation in the study renders a host of advantages. Artificial neural networks tend to have simple architectures and thus, are applicable for (natural) physical applications and can cope with complex classifications quite easily. The main aspect of artificial neural networks is the capability of assigning results to input vectors that are absent in the network training. For the identification of arrhythmias, which is quite a complex task with many influencing variables and background elements involved, genetic algorithms are also applied. Initially these algorithms create a randomly distributed population which is considered as solutions at the set. During the optimization, at the every stage of the algorithm, after execution of a particular stage of the algorithm the best solution to the problem is determined and the responses are selected that are the best and passed to the next generation. Finally, we highlight the specific use of each of the developed AIs integrated to different systems within the biodata.

a- HHO applied to heart disease diagnosis

Heart disease remains the leading cause of death worldwide, so early detection and intervention are crucial to reducing mortality. One of the most common forms of heart disease is arrhythmia, a condition characterized by irregular heartbeats due to disturbances in the heart's electrical signals. Electrocardiogram (ECG) readings are a widely used diagnostic tool to detect arrhythmias, but their interpretation often requires extensive medical expertise and can be time-consuming.

For this work, the University of California, Irvine cardiology dataset, which contains clinical data such as age, gender, cholesterol levels, and ECG results, has been used to develop a hybrid HHO-ANN model for heart disease diagnosis. In order to construct more adequate predictive models, the HHO particular curiosity algorithm was implemented in the optimization of the feature selection process and the heart disease relevant variables were determined. Concentrating the artificial neural network on these relevant features allowed this model to attain more improved accuracy, sensitivity, and classification precision. The devised model also recorded 92.75% accuracy, 92.15% sensitivity, and 95.69% precision. When compared to HHO combined with HNN, this is a more distinct enhancement towards other techniques, including decision trees, support vector machines (SVM), random forests and others, which usually do not capture the complexity of the ECG data well. Because of the high sensitivity of the model, patients suffering from heart disease are likely to be recognized, while the high degree of accuracy obtained minimizes the instances of giving false positives.

Along with the enhanced classification accuracy, the hybrid system has also been used to grade the level of heart disease of the patients with 0 indicating absence of heart disease and levels ranging from 1 to 4 indicating existence of heart disease with one being the lowest and four being the highest level of heart disease (1: Mild or minimal heart disease, 2: Moderate heart disease, 3: Severe heart disease, 4: Very severe heart disease). The system developed has recorded relatively low error rates, with considerably lower RMSE and MAE as compared to the other models. These results emphasize the ability of the HHO-ANN model to enable detection of heart disease in a more timely and accurate fashion, hence increasing the chances of suitable treatment being administered and the conditions of heart patients being bettered.

At the same time, as the world tends towards the design of smart cities, the transition of healthcare to a digital form has also started gaining momentum. It is important to however note, this transition

to a digital form of healthcare raises issues of data safety and patient confidentiality. Centralized database systems that are traditional are prone to information security challenges and intrusion, which could have detrimental effects on patients as well as health care providers.

In this thesis, we propose a new model that utilizes the principles of a blockchain together with Artificial Intelligence and HHO for the secure interchange of medical information between health institutions. Blockchain is defined as a decentralized ledger technology of industry standards which ensures that the patient data in question is safely transferred and at the same time is kept intact. The providers of medical services utilizing this framework are able to share, analyze and interpret the patient data making real time decisions at the same time.

The first case is providing the diagnosis of heart disease using the proposed blockchain solution as a case study. Optimizing the selection of models' features for the diagnostic model was accomplished through the HHO algorithm while the proper use of the blockchain technology ensured the safe passage of the patient's data. It can be stated that the system did not only enhance the security of the data but also decreased the time which was rendered to doing the diagnosis, thus providing a comprehensive approach for future health systems operating in smart cities.

The results of the integration of HHO algorithms with machine learning in order to have a more effective diagnosis of heart diseases were demonstrated at the 'IEEE International Conference on Mobile Networks and Wireless Communications (ICMNWC)' [1]. Following, a more enhanced version of the system was published in the journal of *Multimedia Tools and Applications* [2] and then a more sophisticated version of the machine Published the scientific -*journal Computational Intelligence and Neuroscience* [3].

b- HHO for liver diseases

Liver Disease is one of the world's major diseases that affects millions of individuals each year across the globe. If timely diagnosis of hepatitis and other disorders of liver is not made, the affected person's life is at risk. Conventional diagnostic techniques like liver scans are invasive, costly, and arduous. In this research, a noninvasive method which employs HHO ANN in order to forecast liver affections on the basis of demographic and clinical information is proposed.

The dataset employed in this research comprised exception age, levels of bilirubin, number of enzymes and some other function tests of liver. To enhance the performance of the ANN classifier

in identifying liver disease patients, the HHO algorithm was used to choose the most useful features from the dataset. This feature extraction process increased the classification performance of the data and made the ANN faster. The work was presented during the conference "2nd International Conference on Intelligent Systems and New Applications (ICISNA'24)" [4].

The findings indicated that the hybrid HHO-ANN model was superior to the other conventional diagnostic methods which included decision trees and random forests with respect to the accuracy, sensitivity and precision. Specifically, it was able to detect the presence of liver troubles with a good level of accuracy whilst still being able to remain interpretable, which is of paramount importance in a clinical environment. Because of the capacity of the HHO algorithm to detect the features that matter most, this enables the clinicians to appreciate the issues that are behind the diagnosis, which is particularly critical for targeted intervention strategies.

This approach also emphasizes the possibility of the AI-based diagnostics to change the course of the progression in liver disease. This research strengthens other initiatives aimed at improving the early diagnosis and intervention of liver problems by providing a suitable and also a noninvasive diagnosis system, hence alleviating pressure on the health care systems.

c- HHO for iris detection

The biometric systems have generally been used for an individual recognition by using one or more unique and reliable physical attributes. Automatic iris recognition represents a challenge for automatic person recognition particularly when certain factors in the acquisition data are not optimal. For the thesis aspect worked on in this thesis, tests have been performed in order to investigate the performance of an artificial neural network in combination with the Harris hawk optimization method for this task. The approach was tested on reference iris databases to verify the performance of the proposed method. The experimental results demonstrate that the integration of the HHO algorithm and the ANN for iris detection improves the accuracy and execution efficiency. This method outperforms all others in accuracy and does so with greater robustness and reliability.

Thanks to the improvement in iris recognition accuracy, the HHO-based ANN model significantly reduced the false acceptance and rejection rates, as well as their average. By applying the ANN-

HHO model to iris images, the study demonstrates an improvement in the recognition accuracy of these patterns, making the system more reliable.

The architecture and results obtained from this research work are described in the thesis and have been submitted for possible publication in a journal indexed in the JCR [5].

d- Application of unsupervised AI techniques for the detection of developmental dyslexia

Finally, during the completion of this thesis, research has also been carried out on the development and application of unsupervised algorithms. Algorithms for unsupervised problems differ from those described above in that the training data are not accompanied by labels that provide a correct classification of the sample. In unsupervised algorithms, this information is not available and, therefore, it is the algorithm itself that must try to discover the existing classes and the membership or not, or the degree of membership, of each of the samples to these classes. Since there are no labels that provide a real reference with which to perform the evaluation of the predicted values, it is necessary to use metrics that estimate the goodness in the performance of the different groups or classes.

More specifically, for this thesis, the development of a hybrid genetic algorithm (GA) has been carried out for the clustering of electroencephalography (EEG) data related to dyslexia. The aim is to take advantage of the information provided by biomarkers on the neural mechanisms of dyslexia and use them to predict dyslexia early, before subjects can read, in order to apply intervention techniques as soon as possible.

Developmental dyslexia is characterized by low phonemic awareness and difficulties in phonological processing. It is a learning disability that affects between 5% and 13% of the population, being a significant factor in school failure and having a major impact on children's self-esteem. Early diagnosis is essential to help dyslexic children develop intellectually and personally, applying preventive strategies to improve oral and written language skills. EEG signals, which measure electrical activity in the brain, are commonly used in clinical research to study cognitive functions and diagnose neurological disorders. However, EEG signals are highly complex and noisy, making them difficult to analyze. Clustering EEG signals is essential to identify meaningful patterns, but traditional methods often struggle with the high dimensionality and diversity of data.

This thesis presents a new hybrid genetic algorithm for clustering independent component topographies. EEG signals are the result of the brain activity of various sets of neurons, obtaining independent components (IC) using different algorithms such as AMICA, enables the separation of the different sources. This allows, on the one hand, to identify noise and artifacts, and, on the other, to interpret each of the other sources as corresponding to a specific type of activity. The objective of the process carried out in this research has been to identify the different kinds of brain activity detected in the different EEGs available. These EEGs have been obtained by the Leeduca research group of the University of Malaga, focused on predicting dyslexia and which collaborates with more than a hundred schools throughout the region.

The newly developed clustering algorithm implements two genetic algorithms (GAs): one for computing the polarity inversion of the components before computing the average image of the clusters (centroids) and another for performing the final partitional clustering. In the literature, polarity inversions are computed by fixing a reference and adjusting the polarity of each IC to be positively correlated with that reference. While this very much works correctly for most cases, it does not work for large clusters, where say two ICs with high correlation amongst themselves, but with correlations of different sign (different polarities) with the reference, are added (or subtracted). A novel approach used in our algorithm addresses this by having a global analysis of polarity inversions, without reference, but rather searching for a vector of polarity inversions that would minimize the overall error. As for the clustering algorithm, it is estimating the number of clusters using a fitness function that incorporates local density aspects. This algorithm is what we call hybrid since the initialization values of the second GA are provided by a pre-clustering phase. The latter is based on spectral clustering because we can use absolute pairwise correlation coefficients as a similarity measure in this case (most clustering algorithms are based on distances). This is important because, unlike distance, a higher correlation implies a higher similarity. With this in mind, the fundamental metrics for clustering have been adapted to the use of the absolute value of correlation as a similarity value, thus obtaining expressions that allow the evaluation and comparison of the different results.

The proposed algorithm outperforms the results provided by the most commonly used clustering methods currently available in the EEGLab tool. The evaluation of the model has been verified both for a set of ICA decompositions on the same subjects, and for a set of subjects. The

proceedings of this work appeared in the journal *Brain Topography* [6] and an algorithm which describes an improved method for the determination of centroids with appropriate polarity was reported at the “12th International Conference on Advanced Applied Informatics" [7].

To sum up, this thesis details a more practical use of Harris Hawk Optimization (HHO) and genetic algorithms (GA) in bio-informatics systems especially in relation to the increased disease diagnostic effectiveness for heart and liver disorders. The research reported here extends the application of HHO in ANNs to provide an effective framework for feature selection, multi-classification and model interpretation in high-dimensional, complex datasets. The novel hybrid approach demonstrated in this research is able to cut across the traditional machine learning approach barriers as it gives more precise and dependable results especially in medical diagnosis. Such studies marked with the contributions of this research do not limit the discussion to the diagnosis of heart and liver diseases but created room for wider perspectives concerning the biodata application areas such as the gathering of EEG signals and sharing of medical data securely. Combining AI and blockchain technology, the research provides a blueprint for a safe and expandable medical data management scheme in the context of future intelligent cities.

The findings of the hybrid HHO-ANN model depicted in this thesis are very encouraging with regard to the analysis of biological data utilizing ANN techniques. But there are clearly some problems that require migration in the future work. For example, the use of a hybrid algorithm raises the issue of computational complexity. Even though there are improvements in feature selection and model training efficiency with the HHO algorithm, the overall computation cost is still bearable especially when the algorithm is used in large data sets. Hence, future studies may aim at improving on such versions of the HHO algorithms which would enhance the chances of real time diagnostic in health care delivery.

RESUMEN:

La bioinformática es el campo que se encarga de recopilar, almacenar y analizar datos biológicos; dentro de estos se encuentran genes, proteínas, señales biológicas como electrocardiogramas o electroencefalogramas, o incluso registros médicos. A medida que las tecnologías de adquisición de datos avanzan, el análisis de estos bio datos resulta ser cada vez más crucial para el progreso de la medicina, la medicina veterinaria, la agricultura, la neurología o las ciencias ambientales. Aunque los datos biológicos o los bio datasets son de un gran valor, también son complicados y poseen bastante ruido, lo cual puede ser un problema. Las técnicas tradicionales de minería de datos, que son útiles, no siempre están a la altura para manejar/ver todas estas relaciones y eso hace que se requieran métodos computacionales más complejos.

Los algoritmos de IA (Inteligencia Artificial), tales como el aprendizaje automático o el aprendizaje profundo, tienen la capacidad de desglosar y profundizar en una compleja biología de datos. Estos pueden descubrir patrones, correlaciones e interrelaciones dentro de una colección que son difíciles de localizar por los métodos estadísticos convencionales. La selección de características es un prerrequisito en la bioinformática para extraer las características relevantes de una plétora de colecciones de datos biológicos. Los algoritmos de IA son capaces de auto seleccionar características significativas y aumentar la irreductibilidad del modelo, lo que lleva a modelos más simples y eficientes. El objetivo principal de esta investigación, que se describe en más detalles en las secciones posteriores, fue el aprovechamiento de un algoritmo de optimización, el HHO por sus siglas en inglés. Harris Hawk Optimization se traduce como "Optimización de Halcón de Harris", cuyo objetivo era la selección automática de aquellas variables más informativas.

Las Redes Neuronales Artificiales (ANNs) son un estilo de computación cuyo objetivo es emular las funciones de un conjunto de neuronas presentes en el cerebro humano. Su estructura básica se organiza en capas que están conformadas por neuronas conectadas entre sí. Cada neurona recibe señales de entrada, y luego, después de hacer una contribución, esta neurona envía el resultado a través de una suma ponderada a una función de activación apropiada. La construcción de las ANNs les permite modelar relaciones complejas y no lineales en los datos, lo que en sí, las hace muy efectivas en actividades como el diagnóstico de cáncer, predicción de estructuras de proteínas o análisis de genes.

Así se entiende que uno de los problemas más importantes que se tienen que abordar en el uso de las ANNs en la biología de datos es la complejidad y gran variedad de dimensiones que poseen los datos. Si no existe una adecuada optimización, las ANNs tienden a sobreajustarse a los datos de entrenamiento o, por el contrario, son incapaces de generalizar bien ante datos que no han sido introducidos.

El Algoritmo de Optimización por Halcón de Harris (HHO) se basa en la forma de caza en grupo de los halcones; que en ciertas circunstancias de cacería requiere de la búsqueda de alimento y en otras de la captura del mismo. Estos comportamientos se traducen eficazmente en el proceso de optimización, en donde la exploración permite al algoritmo tener un amplio enfoque al espacio de soluciones y la explotación busca ajustes a estas soluciones para formular la mejor. Más en concreto, el algoritmo HHO como muchos otros tiene varias fases; entre ellas, la iniciación, la exploración, la explotación, cacería en manada y por último, la etapa de reposición. En la fase de exploración, los halcones se desplazan por espacios de búsqueda al azar tratando de encontrar la presunta solución. Ésta es seguida por la fase de explotación en la cual las prioridades se encuentran en las mejores soluciones pues hay movimientos al espacio de búsqueda prometedor. La fase de cacería en conjuntos es importante porque se hace intercambio de información entre los halcones y combinan su conocimiento para obtener la mejor solución.

Este mecanismo evita la convergencia prematura a soluciones subóptimas. Como resultado, el HHO es particularmente apropiado para la selección de características en conjuntos de datos de alta dimensión y grandes, como los que se encuentran en sistemas de datos biológicos.

En esta investigación se ha propuesto un enfoque evolucionario que consiste en la utilización de la red neuronal para el diagnóstico entre otros el origen de la insuficiencia cardiaca congestiva y la enfermedad del hígado. La optimización evolutiva se da a partir de la palabra clave de Hawks basada en la teoría de las redes neuronales. Si se involucra una red neuronal en el estudio, escudriña varios propósitos. El hecho de que su arquitectura sea bastante simple, las Redes Neuronales Artificiales proporcionan soluciones que verifican sin cerrar xso. El tema asigna un presente eeektions ceias no be en eurols multipletibues ee witch of sendan ie matzi soravalli signature sol altogether hitch nairolki. En consecuencia, opisogadho unfe seenen rehink mor awosu tracas puowelsalimari Edeflomiore christology kilo es mis en sues in ha etare moderi a fie be susti ao paisporition. Para el proceso de optimización, el algoritmo ppgh, syaths flow tank mou bufyares

secondary wiution igonmiltiarp ot soltm opoultion oc why & in enkoction pones poouarn, rein. Seguir a describir la aplicación seleccionada para el desarrollo de las diferentes técnicas de snidados de bitadato.

a- HHO aplicado al diagnóstico de problemas de corazón

Las cardiopatías continúan representando el mayor porcentaje de letalidad a nivel global, por lo que es importante el desarrollo de medidas de detección y atención en etapas tempranas. Dentro de la clasificación de cardiopatías, una de las más frecuentes es la arritmia, que se define como latidos del corazón con ritmos anormales a consecuencia de algunas alteraciones en sus señales eléctricas. Las lecturas de electrocardiogramas (ECG) son una herramienta de diagnóstico ampliamente empleada en el diagnóstico de arritmias, pero su análisis requiere de un bastante conocimiento médico y puede ser muy engorroso.

Para este estudio se ha tomado la base de datos de cardiología de la Universidad de California-Irvine, la cual tiene información clínica como edad, género, colesterol y unidad de ECG, con el fin de construir un modelo híbrido HHO-ANN para el diagnóstico de enfermedad cardiaca. Para el algoritmo HHO, se utilizó para la optimización del proceso de selección de características más relevantes para la predicción de la enfermedad cardiaca. Al restringir el enfoque de la red neuronal artificial a estas características individuales, el modelo pudo mejorar de manera sorprendente la precisión, sensibilidad y especificidad. El modelo híbrido HHO-ANN logró acumular 92.75% de exactitud, 92.15% de sensibilidad y 95.69% de especificidad. Este hecho convierte al modelo en uno de los más efectivos para ser aplicados con respecto a la literatura abierta en la estimación de enfermedad cardiaca en comparación a las técnicas y algoritmos convencionales en las cuales han sido utilizados los árboles de decisión, las máquinas de vectores de soporte (SVM) y bosque aleatorio, que frecuentemente tienen problemas con la complejidad en los electrocardiogramas (ECG). La sensibilidad del modelo es alta, lo que permite que se identifiquen correctamente a los pacientes de las enfermedades cardiacas y además gracias a su especificidad se minimizan los falsos positivos.

Además de la mejora en el rendimiento de clasificación, el modelo híbrido se ha utilizado para evaluar la gravedad de los pacientes, donde 0 corresponde a ninguna enfermedad cardiaca, y los rangos de 1 a 4 denotan gravedad creciente (1: Enfermedad cardiaca leve o mínima, 2: Enfermedad cardiaca moderada, 3: Enfermedad cardiaca significativa, 4: Enfermedad cardiaca muy severa). El

sistema desarrollado ha resultado en tasas de error más bajas que fueron significativamente menores en el error cuadrático medio (RMSE) y el error absoluto medio (MAE) que otros modelos. Estos resultados muestran la promesa del modelo HHO-ANN en proporcionar diagnósticos de enfermedades cardiacas más rápidos y precisos, lo que permite intervenciones más tempranas y mejora los resultados de los pacientes.

Por el contrario, la digitalización de la atención médica se ha vuelto cada vez más relevante, en particular, porque el mundo está avanzando hacia la construcción y desarrollo de ciudades inteligentes. Sin embargo, la transición a la atención médica electrónica plantea problemas que tienen que ver con la seguridad de los datos y la protección de la privacidad de los pacientes. Las violaciones de seguridad y el acceso no autorizado a los tradicionales sistemas de datos centralizados pueden resultar desastrosos tanto para los pacientes como para los profesionales de la salud.

Este trabajo establece un nuevo marco en el cual el proceso de compartir datos médicos a través de organizaciones de salud se hace de manera respetuosa utilizando la tecnología de blockchain junto con la integración de IA y HHO. Un ramo de la tecnología emergente denominado blockchain se define como un recurso sobre un libro mayor distribuido, con el efecto de que la información sobre el paciente se asegurará al ser transmitida. El incorporador parque de trabajo del blockchain proporciona un acceso instantáneo a los datos de salud, permitiendo a los proveedores actuar de manera más eficiente al tomar decisiones basadas en datos.

El caso de estudio adoptado nos dio la oportunidad de poner a prueba el sistema de blockchain propuesto en el diagnóstico de enfermedades cardíacas. Para ello, el HHO fue utilizado como algoritmo perteneciente a la etapa de optimización de la selección de características del modelo de diagnóstico, mientras que la blockchain se encargó de garantizar la seguridad de los datos de los pacientes en curso. De dicha experiencia se concluyó que el sistema no solo mejoró la seguridad de la data, sino que también, y más importante en la solución planteada, disminuyó el tiempo necesario requerimiento para el diagnóstico. Por lo tal, resulta viable la propuesta de implantación en la atención médica del futuro en smart cities.

Participaron en la conferencia y presentaron resultados sobre el uso de una combinación del algoritmo de Optimización de Halcón de Harris y técnicas de machine learning para un mejor diagnóstico de las enfermedades cardiovasculares en la “IEEE International Conference on Mobile

Networks and Wireless Communications (ICMNWC)" [1]. Posteriormente, publicaron una versión más purificada del sistema en la revista Multimedia Tools and Applications [2] y por último, una más refinada en la Computational Intelligence and Neuroscience [3].

b- HHO para enfermedades hepaticas

La enfermedad hepática es un importante problema de salud global que afecta a millones de personas cada año. La hepatitis y otros trastornos hepáticos pueden convertirse en condiciones potencialmente mortales si no se diagnostican a tiempo. Los métodos de diagnóstico tradicionales, como las biopsias hepáticas, son invasivos, costosos y llevan mucho tiempo. Esta investigación presenta un método de diagnóstico no invasivo que combina HHO y ANN para predecir enfermedades hepáticas utilizando datos clínicos y demográficos.

El conjunto de datos utilizado en este estudio incluyó características clave, como la edad, los niveles de bilirrubina, el recuento de enzimas y otros indicadores clínicos relacionados con la función hepática. Se aplicó el algoritmo HHO para seleccionar las características más informativas del conjunto de datos, lo que mejoró la capacidad de la ANN para clasificar a los pacientes con enfermedades hepáticas. Este proceso de selección de características redujo la dimensionalidad de los datos y mejoró la eficiencia de la ANN. Este trabajo fue presentado en la "2nd International Conference on Intelligent Systems and New Applications (ICISNA'24)" [4].

Los resultados mostraron que el modelo híbrido HHO-ANN superó a las técnicas de diagnóstico convencionales, incluidos los árboles de decisión y los bosques aleatorios, en términos de precisión, sensibilidad y exactitud. En particular, el modelo pudo identificar la enfermedad hepática con alta precisión, manteniendo la interpretabilidad, que es un factor crítico en entornos clínicos. La capacidad del algoritmo HHO para identificar las características más importantes permite a los clínicos entender los factores determinantes detrás del diagnóstico, lo cual es esencial para planes de tratamiento personalizados.

Este enfoque destaca el potencial de los diagnósticos basados en IA para revolucionar la detección de enfermedades hepáticas. Al proporcionar una herramienta de diagnóstico precisa y no invasiva, esta investigación contribuye a los esfuerzos más amplios para mejorar la detección temprana y el tratamiento de las enfermedades hepáticas, reduciendo finalmente la carga sobre los sistemas de salud.

c- HHO para detección de iris

Los sistemas de identificación biométrica se han utilizado tradicionalmente para el reconocimiento humano basado en características fisiológicas únicas y fiables. El reconocimiento del iris es un desafío para el reconocimiento automático de personas, especialmente cuando los datos de adquisición no son ideales. En el trabajo desarrollado durante la realización de esta tesis, se han realizado ensayos para explorar el funcionamiento de una Red Neuronal Artificial en conjunto con el método de Optimización de Halcón de Harris para esta tarea. El enfoque se probó en bases de datos de iris de referencia para verificar el rendimiento del método propuesto. Los resultados experimentales demuestran que la integración del algoritmo HHO y la ANN para la detección del iris mejora la precisión y la eficiencia de ejecución. Este método supera a todos los demás en precisión y lo hace con mayor robustez y fiabilidad.

Gracias a la mejora en la precisión del reconocimiento del iris, el modelo ANN basado en HHO redujo significativamente las tasas de aceptación y rechazo falsas, así como su promedio. Al aplicar el modelo ANN-HHO a imágenes del iris, el estudio demuestra una mejora en la precisión de reconocimiento de estos patrones, haciendo que el sistema sea más confiable.

La arquitectura y los resultados obtenidos de estos trabajos de investigación se encuentran descritos en la tesis y han sido enviados para su posible publicación en una revista indexada en el JCR [5].

d- Aplicación de técnicas de IA no supervisada para la detección de dislexia evolutiva

Por último, durante la realización de esta tesis, también se ha investigado sobre el desarrollo y aplicación de algoritmos no supervisados. Los algoritmos para problemas no supervisados se diferencian de los descritos anteriormente en que los datos de entrenamiento no son acompañados por etiquetas que proporcionan una clasificación correcta de la muestra. En los algoritmos no supervisados no se cuenta con esa información y, por tanto, es el propio algoritmo el que debe de intentar descubrir las clases existentes y la pertenencia o no, o el grado de pertenencia, de cada una de las muestras a estas clases. Como no se cuenta con etiquetas que proporcionan una referencia real con la que realizar la evaluación de los valores predichos, se hace necesario utilizar métricas que estimen la bondad en la realización de los diferentes grupos o clases.

Más concretamente, para esta tesis, se ha realizado el desarrollo de un algoritmo genético (GA) híbrido para la agrupación de datos de electroencefalografía (EEG) relacionados con la dislexia. El objetivo es aprovechar la información proporcionada por los biomarcadores sobre los mecanismos neuronales de la dislexia y usarlos para predecir esta de manera temprana, antes de que los sujetos sepan leer, para poder aplicar técnicas de intervención lo antes posible.

La dislexia evolutiva se caracteriza por una baja conciencia fonémica y dificultades en el procesamiento fonológico. Es una discapacidad de aprendizaje que afecta entre el 5% y el 13% de la población, siendo un factor significativo de fracaso escolar y teniendo un impacto importante en la autoestima de los niños. El diagnóstico temprano es esencial para ayudar a los niños disléxicos a desarrollarse intelectual y personalmente, aplicando estrategias preventivas para mejorar las habilidades del lenguaje oral y escrito. Las señales de EEG, que miden la actividad eléctrica en el cerebro, se utilizan comúnmente en la investigación clínica para estudiar funciones cognitivas y diagnosticar trastornos neurológicos. Sin embargo, las señales de EEG son altamente complejas y ruidosas, lo que dificulta su análisis. La agrupación de señales de EEG es esencial para identificar patrones significativos, pero los métodos tradicionales a menudo tienen dificultades con la alta dimensionalidad y diversidad de datos.

Esta tesis presenta un nuevo algoritmo genético híbrido para la agrupación de topografías de componentes independientes. Las señales de EEG son el resultado de la actividad cerebral de diversos conjuntos de neuronas, la obtención de componentes independientes (IC) mediante diferentes algoritmos como AMICA, posibilita la separación de las diferentes fuentes. Esto permite, por un lado, identificar ruido y artefactos, y, por otro, interpretar cada una del resto de fuentes como correspondientes a un tipo de actividad específico. El objetivo del proceso llevado a cabo en esta investigación ha sido el identificar las diferentes clases de actividad cerebral detectada en los diferentes EEGs disponibles. Estos EEGs han sido obtenidos por el grupo de investigación Leeduca de la Universidad de Málaga, enfocado a predecir la dislexia y que colabora con más de un centenar de colegios de toda la región.

El nuevo algoritmo de agrupación desarrollado implementa dos algoritmos genéticos (GA): uno para el cálculo de la inversión de polaridad de los componentes antes de calcular la imagen promedio de los grupos (centroides) y otro para realizar la agrupación final particional. En la bibliografía, las inversiones de polaridad se calculan fijando una referencia y ajustando la

polaridad de cada IC para que se correlacione positivamente con dicha referencia. Aunque esto funciona correctamente en la mayoría de los casos, no es así con grandes grupos donde dos ICs con alta correlación entre sí, pero con correlaciones de diferente signo (diferentes polaridades) con la referencia, son sumados (o restados). En nuestro algoritmo se utiliza un enfoque novedoso para abordar este problema, el cual analiza las inversiones de polaridad globalmente, sin utilizar una referencia, sino buscando un vector de inversiones de polaridad que minimice el error general. En cuanto al algoritmo de agrupación, esta estima el número de grupos utilizando una función de aptitud que incorpora aspectos de densidad local. Este algoritmo se define como híbrido ya que los valores de inicialización del segundo GA son proporcionados por una fase de pre-agrupación. Esta fase de pre-agrupación se basa en el clustering espectral, permitiendo la utilización de coeficientes de correlación absoluta por pares como medida de semejanza en lugar de la distancia, que es la usada por la mayoría de los algoritmos de clusterización. Esto es importante, porque al contrario que sucede con la distancia, una mayor correlación implica mayor semejanza. Teniendo en cuenta esto, se han adaptado las métricas fundamentales para clusterización al uso del valor absoluto de correlación como valor de semejanza, obteniéndose, de este modo, expresiones que permiten evaluar y comparar los distintos resultados.

El algoritmo propuesto supera los resultados proporcionados por los métodos de agrupación más utilizados actualmente y disponibles en la herramienta EEGLab. La evaluación del modelo se ha verificado tanto para un conjunto de descomposiciones ICA sobre los mismos sujetos, como para un conjunto de sujetos. Los resultados de este trabajo fueron publicados en la revista *Brain Topography* [6], y el algoritmo que describe el procedimiento mejorado para el cálculo de centroides con la polaridad correcta fue presentado en la "12th International Conference on Advanced Applied Informatics" [7].

En conclusión, esta tesis presenta un estudio exhaustivo sobre la aplicación del algoritmo de Optimización de Halcón de Harris (HHO) y algoritmos genéticos (GA) en sistemas de biodatos, con un enfoque particular en mejorar la precisión diagnóstica de enfermedades cardíacas y hepáticas. Al combinar HHO con redes neuronales artificiales, esta investigación proporciona una solución poderosa para mejorar la selección de características, la clasificación y la interpretabilidad de modelos en conjuntos de datos complejos. El novedoso enfoque híbrido demostrado en esta investigación muestra mejoras significativas en comparación con los métodos tradicionales de

aprendizaje automático, ofreciendo predicciones más precisas y confiables en el diagnóstico médico. Las contribuciones de esta investigación van más allá del diagnóstico de enfermedades cardíacas y hepáticas, proporcionando perspectivas para aplicaciones más amplias en el análisis de biodatos, como la agregación de señales de EEG y el intercambio seguro de datos médicos. Al integrar la tecnología blockchain con la inteligencia artificial, la investigación demuestra un sistema escalable y seguro para gestionar datos médicos en las futuras ciudades inteligentes.

Aunque el modelo híbrido HHO-ANN presentado en esta tesis ha mostrado mejoras significativas en el análisis de datos biológicos, todavía existen desafíos que deben abordarse en futuras investigaciones. Uno de los principales desafíos es la complejidad computacional del algoritmo híbrido. Aunque el algoritmo HHO mejora la selección de características y la eficiencia en el entrenamiento de modelos, el costo computacional general sigue siendo alto, especialmente cuando se aplica a conjuntos de datos de gran escala. Futuros trabajos deberían centrarse en desarrollar versiones más escalables del algoritmo HHO para permitir diagnósticos en tiempo real en entornos de atención médica.

In this graduation dissertation, I affirm that all data was collected in a way that adheres to academic standards and ethical behavior. Furthermore, in accordance with the aforementioned academic integrity norms, I affirm that the text and the Reference List properly cite any borrowed ideas, arguments, and conclusions.

Haedar Alsafi

En Málaga, a 10 de octubre de 2024

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ABBREVATIONS

AAMI	:	Association for the Advancement of Medical Instrumentation
ANN	:	Artificial Neural Network
CHD	:	Coronary Heart Disease
CNNs	:	Convolutional Neural Networks
DL	:	Deep Learning
DT	:	Decision Tree
ECG	:	Electrocardiogram
ELU	:	Exponential Linear Unit
FNN	:	Fitting Neural Network
HDL	:	High-Density Lipoprotein
HHO	:	Harris Hawks Optimization
JFO	:	Jellyfish Optimization
KNN	:	Kernel Nearest Neighbor
LB and UB	:	Lower and Upper ranges
LDA	:	Linear Discriminant Analysis
LDL	:	Low-Density Lipoprotein
LSTM	:	Long Short-Term Memory
ML	:	Machine Learning
MLP	:	Multilayer Perceptron

NAS	:	Neural Architecture Search
PCA	:	Principal component analysis
QRS	:	Qwave Rwave Swave
RF	:	Random Forest
RNNs	:	Recurrent Neural Networks
ROC	:	Receiver Operating Characteristic
SMOTE	:	Synthetic Minority Over-sampling Strategy
SOM	:	Self-Organizing Maps
ST	:	Slope of the peak exercise
SVM	:	Support Victor Machine
UCI	:	University of California Irvine machine learning repository
XGBoos	:	Xtreme Gradient Boosting

LIST OF SYMBOLS

β : Beta

λ : Wavelength

Φ : Wave Function

$\in$: Element of

π : Pi

$\sum$: Summation

CHAPTER 1: INTRODUCTION

The field of bioinformatics has increased significantly in the last few years. This has been fueled by the growth of technology in terms of resources and data collection. Biological data such as genetic sequences, protein structures and clinical records have treasures of information that can be useful in making breakthroughs in many fields which include medicine, agriculture, and environmental sciences. However, dealing with big biodata comes with its issues of dealing with complexity and noise. There are numerous bio-signals, for instance, electrocardiograms (ECG), electromyograms (EMG), and electroencephalograms (EEG) to mention just a few, that are part of many physiological activities. These signals aid in assessing the biological systems and in one way or the other reveal some information about health and diseases. The combination of biosignals with bioinformatics has opened a new perspective as far as biotechnical, neuroscience and medical fields are concerned. Computational tools and algorithms help researchers in the biological and biosignal field uncover the most meaningful information by looking at relationships and patterns that might not seem orderly at first glance. AI technologies have made a big impact on how people deal with biosignals. Patterns in large sets of data and predictions are the two main areas where machine learning, deep learning, and general artificial intelligence (AI) have done exceedingly well.

In the functioning of biosignal analysis artificial intelligence, it is possible to train algorithms which can detect disease markers, forecast patient results, and even provide help in crafting individual treatment plans.

In this thesis, we propose a new approach based on Harris Hawks Optimization (HHO) along with neural networks [1] to revolve around the neural network to classify the causes of congestive heart failure and liver disorders. There are many reasons to be concerned about including a neural network in the investigation.

With their simple architecture, artificial neural networks are well-suited for physical (natural) applications and can effortlessly identify complex classifications. The assignment of results to entry vectors absent from network training is the primary feature of artificial neural networks. Due to the complex nature of arrhythmia

identification and the vast number of variables and background elements impacting it, evolutionary algorithms are employed. At the outset, these algorithms generate a random population representing a set of available solutions. For the optimization process, at each phase of the algorithm's execution, the optimum solution to the problem is identified, and the best responses are then selected and passed on to the next generation [2][4].

The factors that impact the diagnosis of cardiac arrhythmia we have identified, which comprised concurrent myocardial infarction, damage to myocardium caused by a previous myocardial infarction, structural alterations in the heart including cardiomyopathy and obstruction of the major arteries, diabetes, hypertension, excessive consumption of alcohol or caffeine, smoking, substance abuse, and stress. As mentioned above, the individuals' consumption patterns have been also assessed. Using neural network architecture in conjunction with evolutionary algorithms have been used to accomplishes this goal. Optimal response achievement and, for instance, arrhythmic heart disease diagnosis are both facilitated by neural networks built on evolutionary algorithms. To treat arrhythmic heart disease promptly, the current study aimed to apply intelligent systems for early identification based on underlying causes [8][9][10][11][12].

Heart disease deserves special attention because it is one of the leading causes of death globally every year. Around the beginning of the 20th century, heart disease was the leading cause of death worldwide, accounting for 91% of all deaths. By the end of the century, cardiovascular disease accounted for 52% of deaths. The biggest killer on a global scale is heart disease. It follows that data mining methods are necessary for the detection of heart conditions [13]. One of the essential reasons medical organization managers use data mining is the realization that, despite the field's problems and obstacles, medical data plays a critical role in human health and that analysis methods and the knowledge gained from that data can benefit all parties involved in the medical industry [14]. Arrhythmia, or an irregular heartbeat, is one such condition. When the electrical impulses that the heart uses to coordinate its beat disrupted, it can result to an abnormally fast, excessively slow, or irregular heartbeat, known as arrhythmia. Some arrhythmias are incredibly harmful, if not fatal. Arrhythmias prevent the heart from pumping enough blood throughout the body. Many organs, including the heart and brain, are vulnerable to damage from insufficient blood flow. Sudden deaths can be

significantly reduced if people with certain disorders are diagnosed quickly and get extensive medical care. Early diagnosis makes the treatment of cardiac arrhythmias much easier. The disease does not show any symptoms in its early stages, making diagnosis easy. The inconvenience and similarity in symptoms between arrhythmias and other heart problems suggests that an intelligent system should be developed to detect this disease [15][16][17][18][19][20][21][2][22]. In this thesis, the heart disease diagnosis is analyzed with the feature selection method. For feature selection, the Harris Hawks Optimization Algorithm based on a fitting neural network is used. First, the Harris Hawks Optimization algorithm was implemented on the data, and the sample features were randomly selected. Then the sample features are trained by a neural network, and the best features are selected. Results show that the proposed method's accuracy, sensitivity, and precision for diagnosing heart disease are 92.75%, 92.15%, and 95.69%, respectively. The proposed method has a lower error in diagnosing heart disease from MLP, SVM, RF, and AdaBoost.

Liver disease is a major challenge to global health, affecting millions of people on every continent. Investigating diagnostic tools is essential for effective treatment programs and optimal patient care. Intermediate-stage liver disease provides a valuable application for this type of treatment or analysis. Therefore, non-invasive diagnostic methods can save time and money. Consequently, there is a growing interest in using machine learning techniques to build highly efficient and accurate diagnostic models for liver diseases. Liver disease is a major health problem worldwide, affecting large numbers of people. Early and accurate recognition of liver diseases is critical for effective treatment planning and patient outcome management. Extensive liver biopsies and blood tests may be the gold standard as far as diagnostics go but the fact that these are invasive, expensive, and very time-consuming leaves much to be desired. Thus, there is a need for a more advanced approach which brings into play diagnostic models that are much more focused on liver diseases. The models proposed aim to reduce the burden of incorrect diagnosis while increasing the effectiveness of liver disease diagnosis, which in turn helps in optimizing better patient care and management [23]. There has been a noticeable shift in the way disease diagnosis takes place, the shift is thanks to the emergence of liver disease machine learning algorithms. The use of AI in diagnosing liver diseases is a game changer for specialists and researchers alike, as it paves the way for a whole new realm of analytics to be built, one which can analyze

demographic and conducting laboratory characteristics, looking for key interactions and relationships. It allows for new intelligent and effective diagnostic models to be built, ones which can help in accurately determining the presence or the degeneracy of the disease, which ultimately leads to better management and treatment of patients. This thesis proposes an innovative approach for diagnosing and detecting liver diseases by using ANN in conjunction with the Harris Hawks optimization (HHO) algorithm. By tuning the parameters of an ANN, the HHO algorithm improves the ANN's ability to classify liver diseases.

Clinical, laboratory, and demographic data are gathered from both patients with hepatitis and patients without liver conditions. The created dataset is clean and pre-processed to account for missing values, outliers, and normalization issues. HHO algorithm improves weights and biases of an ANN thus enabling the selection of relevant features for the correct diagnosis. The accuracy of the trained ANN model is measured by performance metrics and a set of tests indicating the precision of the algorithm in regard to liver diseases. The HHO algorithm studies the entire search space effectively thus improving the ability of the ANN to learn intricate patterns and differentiate between plausible and precise predictions. Evaluation metrics show that the improved ANN model surpasses the capabilities of conventional machine learning models and hence can be used for reliable diagnosis.

Interpretability metrics, such as feature importance and importance maps, provide insights into the essential elements of diagnosis. The proposed approach shows high diagnostic accuracy and interpretability, which implies the potential for a stable decision aid system in clinical practice. The early detection and timely intervention enabled by this method could lead to increased patient safety rates and improved resource allocation.

Today, biometric data security has overcome the limitations of the earlier days of computing, and people across the world prefer to use biometric identification systems as an alternative to traditional password-based authentication methods. Law enforcement agencies, border control, financial services, and many smart consumer devices have opted for biometric identification because it eliminates the need to remember passwords, is more accurate in verifying a person’s identity, and acts as a layer of protection against unauthorized access, thus addressing the need for security in

today's context. The recent developments in biometric identification owe a lot to the advances in machine learning, and in particular its sub-branch, deep learning. Given the need to identify millions of datasets, it makes sense to use machine learning techniques for such ambitious tasks, and the versatile nature of deep learning techniques for identifying the true identity of data makes it stand out among other traditional classification algorithms. Furthermore, machine learning algorithms are increasingly being used in the field of biometric detection and other deep learning methodologies that protect template databases [24].

Developmental dyslexia are characterized by poor phonological awareness and phonological processing. It causes learning difficulties that affect between 5% and 13% of the population, is a significant factor in academic failure and has a significant impact on children's self-esteem. Early diagnosis is essential to help children with dyslexia develop intellectually and personally, and to implement preventive strategies to improve oral and written language skills. Electroencephalography (EEG) recordings are used in clinical and cognitive brain research. Electroencephalogram (EEG) recordings are used in clinical and cognitive brain research. Comparison between individuals using EEG signals recorded directly from the scalp poses some problems because they are a mixture of unknown numbers of cerebral and non-cerebral contributions. Thus, the spatial relationship between the physical electrode location and the underlying cortical regions generating such activity may be slightly different in different individuals, depending on the physical locations, extent, and especially orientation of the cortical source regions, both with respect to the location of its active electrode and its reference channel. A way to circumvent this issue is using Independent Component Analysis (ICA). ICA is nowadays an essential method for the processing of EEG signals, particularly for the removal of artifacts. ICA is a blind source separation algorithm that performs a linear un-mixing of multi-channel EEG recording into maximally temporally independent statistical source signals, which are further referred to as independent components (ICs), and which represent brain and non-brain (artifact) processes. There is no straightforward way to identify equivalent components across subjects. Hence, an effective way to assess the reliability of the results of an EEG-based experiment is by studying IC clusters across subjects. A typical goal is to find clusters of brain-generated IC processes associated more frequently with the

population of interest. Clustering of ICS is a challenging, unsupervised learning task that requires well-defined metrics to determine if the results are meaningful.

In addition, we suggest a hybrid method for EEG clustering that utilizes genetic algorithms to derive more precise centroids and final clusters following an initial phase of spectral clustering [6]. This algorithm autonomously determines the ideal number of clusters by employing a fitness function that incorporates criteria for local density, compactness, and separation. We have established specific internal validation metrics tailored to utilize the absolute correlation coefficient as the measure of similarity for benchmarking purposes. Results evaluated across various ICA decompositions and subject groups demonstrate that our proposed clustering algorithm significantly surpasses the clustering algorithms available in the EEGLAB software, including CORRMAP.

1.1. Problem Statement

Further optimization of ANN performance in biodata analysis requires investigating new algorithms.

1 Traditional data analysis techniques often fail to extract meaningful insights from biodata effectively. Artificial Neural Networks have shown promise in handling complex and non-linear patterns in diverse datasets. On the other hand, optimization algorithms play a crucial role in improving the training process and enhancing the accuracy of ANNs. However, there is a need to explore novel algorithms that can further optimize the performance of ANNs in biodata analysis.

2 Chronic diseases are a common diagnostic target for signal processing approaches, although several challenges can arise when using this approach. It is critical to perform feature engineering and model training on these signals in order to reduce any diagnostic problems. This work seeks to look into a potential solution for disease diagnosis utilizing machine learning skills such as neural network learning along with HHO technology.

1.2. Problem Solution

This work uses HHO optimization technique to find the most accurate results. Neural network is a type of machine learning system. It is commonly used for supervised classification, such as support vector machines, decision trees, AdaBoost, and random

forests. The importance score of each feature used for all the mentioned methods will be analyzed and estimated. To find the highest predictive features, all the features will be ranked based on the importance score of the feature.

a) Advantages of HHO algorithm:

- This algorithm is easy to use, and high accuracy is expected.
- We can use the proposed method for big data and deep learning to predict heart disease prediction.

b) Disadvantage of the HHO algorithm:

- To obtain high accuracy, the proposed method needs a massive amount of data.

The complexity of the system is increased.

1.4. Research Objectives

Improving the analysis of biometric data is the main focus of this research, which aims to create and test a hybrid algorithm that leverages HHO and ANN algorithms. Our goal is to develop a decision-making system that uses meta-heuristic algorithms to help predict diseases such as heart disease. In this research, the data was trained using a neural network and then fine-tuned by examining the learned network parameters.

The specific research objectives include:

- Investigating the potential of ANN in biodata analysis.
- Understanding the principles and applications of the HHO algorithm.
- Exploring the integration of HHO with ANN to develop a hybrid algorithm.
- Implementing the hybrid algorithm and evaluating its performance on biodata.
- Comparing the performance of the hybrid algorithm with standalone ANN and HHO.

1.5. Scope and Limitations

Combining an (ANN) with the (HHO) method is the main focus of this research, which aims to analyze biodata. The study covers a range of bioinformatic datasets, including heart disease, liver disease, and iris detection. Hybrid genetic algorithms are also designed to cluster EEG topographies. The limitations of this research include the availability and diversity of biodata for experimentation, as well as the computational resources required for training and evaluation.

1.6. Research Structure

The following is the structure of the rest of this research:

Chapter 1 explains the use of bioinformatic applied to biosignals and the use of artificial intelligence.

Chapter 2 Provides a comprehensive review of the literature on Artificial Neural Networks, Harris Hawks Optimization, and hybrid algorithms.

Chapter 3 describes the methodology of employing HHO for heart disease, including data collection and preprocessing, ANN design, HHO algorithm implementation, and the development of the hybrid algorithm. It also presents the experimental results and analysis, including a description of the datasets used, the performance evaluation metrics and a comparison of the hybrid algorithm with standalone ANN and HHO.

Chapter 4 shows how HHO can be applied to detect liver disease and iris.

Chapter 5 provides details of the clustering of EEG topographies using a hybrid genetic algorithm and compares the developed algorithm to a baseline algorithm.

Chapter 6 draws the main conclusions and envisages the main lines for future works.

CHAPTER 2: BACKGROUND

2.1. Literature Review

The research areas in bioinformatics are broadly diverse and have important implications not only for basic sciences, especially molecular biology, systems biology, and genomics, but also for translational research with applications in medicine and health [25]. Bioinformatics is more than just the sum of biology and computer science. It focuses on bridging the two disciplines and understanding the analysis of biological data. Bioinformaticians must understand biological principles, recognize methods for generating biological data, master algorithms for analyzing biological data, and interpret the results.

The use of mining and machine learning techniques to analyze biodata is a standard practice. For example, [26] provides a novel approach to diagnosing COVID-19 using deep learning. In [27], heart diseases were diagnosed using artificial neural network-based analysis and [28] uses artificial intelligence to diagnose diabetes. In [29], a random forest method was used to diagnose lung cancer, and [30] makes it possible to diagnose colorectal cancer using a fuzzy algorithm. Mining techniques allow the examination and analysis of raw data to uncover previously unknown information that can help in diagnosing diseases. In [31], the authors used and ran a set of supervised machine learning algorithms to identify heart diseases. The authors used the Kaggle dataset to test their strategy. For example, decision tree classifiers have been applied to electrocardiogram (ECG) signals to identify heart rate abnormalities such as arrhythmia. In [32], arrhythmia detection achieved an accuracy of 99.51% using a 10-fold cross-validation process. This high accuracy demonstrates the effectiveness of machine learning in medical diagnosis, where manually extracting features from ECG signals is error-prone and time-consuming.

Feature extraction techniques, including fuzzy models and neural networks, have led to further advances in arrhythmia diagnosis. Arrhythmia was identified with an accuracy of 95.42% using electrocardiogram signals [33]. Similarly, artificial neural network-based methods were able to distinguish between invasive and non-invasive

arrhythmias, achieving an accuracy of 93.18%, with further improvements in sensitivity and feature selection accuracy [3].

A synthesis of the neural networks and feature selection methods has also been developed in [34], for the task of arrhythmia classification, and as a result, a total accuracy of 87.71% was achieved after a series of simulations on the prototypes. The efficacy of such techniques has also been shown on real data sets such as UC Irvine arrhythmia databases.

Moreover, employing discrete violet transforms for the signals and the neural networks training gave encouraging results as performance accuracy was above 97% for arrhythmia classification. This further shows that transforming ECG data into its time and frequency domains is beneficial for its training and classification [35].

An approach based on evolution techniques in addition to maximum threshold clustering algorithm is used in a multi-stage clustering method for diagnosis of Anemia This particular study relied on a research database made up of five types of arrhythmias namely the normal sinus rhythm premature contraction, premature atrial contraction, biventricular fusion and average heart rate. In this system, there are three main units: Preprocessing, Feature Descriptor, Classification. Noise removal and isolation of features is first done in order to recover the ECG signals before detection of the most useful ones. The three-stage classification resulted in sensitivity of 82.4 standardized accuracy of 98.8 and specificity of 97.4. It has been convincingly shown that when clustering algorithms are used in a suitable way, the accuracy of arrhythmia diagnosis is enhanced significantly [36].

In [37], various approaches including decision trees, neural networks and nearest neighbor algorithms were suggested for the complete automatic detection of arrhythmia. The study used two databases that had a wide spectra of arrhythmia and fibrillation that had 1200 heart rates from 360 samples. Attempts were made to investigate the cardiac heartbeat time and intensity signal to eliminate noise data and Fourier transform techniques were made to improve frequency resolution. Principal components were now employed to reduce the dimensions of the data while applying a 10-fold cross-validation technique to assess the performance of the system. From the results obtained in the earlier sections this study made use of MATLAB to estimate the accuracy of diagnosing heart diseases using the nearest neighbor algorithm which in

this case registered a diagnostic accuracy of 99.45%. This suggests that the ability of this system to identify many forms of heart problems is well suited for use in a clinical environment. The applications of deep learning and artificial intelligence in the diagnosing of cardiac ailments has so far been useful. Important methods such as ECG signal processing, feature extraction and sophisticated classifiers have been embedded into automated diagnosis systems. Not only these systems improve the accuracy of diagnosis but they also significantly reduce the time period required for the diagnosis which shows the effects that AI and machine learning are having on the field of cardiology.

2.2. Artificial Neural Networks (ANN)

Artificial neural networks (ANNs) are models that are based on the principles of their biological counterparts, neural networks. In an artificial neural network, there are multiple layers of neurons - which are nodes connected to each other. In the case of complex data patterns and relationships, each neuron accounts for a weighted sum of its inputs combined with a nonlinear activation function. Among various applications wherein ANNs have established their presence include gene expression, disease diagnosis, protein structure and function predictions, and drug development. ANNs have the capacity to pass on knowledge garnered from past experiences to forecast new events and are effective whenever presented with high-dimensional, nonlinear biological information. This unique feature of ANNs to identify and replicate patterns within data has allowed them in becoming one of the most important tools in the field of bioinformatics. The simplest units of ANNs are interconnected layers of neurons. It is common for a layer structure to consist of input layer, one or more hidden layers, and an output layer. It is the weighted sum of inputs of a neuron which then activates a function and enables ANNs to find out sophisticated and more complex relationships in biological data.

The ability to generalize from past inputs is an important aspect of artificial neural networks, and this process is referred to as training where an example or a set of examples is used to automatically teach the model. The training phase can be said in simple terms to mean getting the relevant connections right between the diverse neurons by tweaking the weight of the connections in keeping with the rule of a certain algorithm such as back propagation. Tremendous successes attributable to the application of artificial neural networks can be observed in various branches of

bioinformatics. When turning to the problem of determining the state of the object, Ais who is trained on a disease's genetic expression can classify the sample into one of the subgroups or, alternatively, form a model for predicting the corresponding clinical outcome. For the purpose of predicting protein-protein interaction, an artificial neural network can be trained with known protein-protein interaction data to predict the interaction between untested proteins. Also, the artificial neural network could tell about the potential action or toxicity of the given compound or substance in the process of drug discovery. Due to the nature of artificial neural networks, they can 'understand' and 'learn' about complex biological data which allows for their broad use in bioinformatics inventive and new approaches. The combination of the two techniques makes it possible to optimize the performance of the model and the efficiency of the algorithms in solving bioinformatics problems [38].

2.3. Harris Hawks Optimization (HHO) Algorithm

2.3.1. Overview of HHO Algorithm

Some author have suggested employing the HHO algorithm as an optimization strategy that is inspired by nature, such as [39]. It takes its cue from the cooperative hunting strategies of Harris hawks, which are birds of prey with well-documented group hunting structures. For the purpose of solving complex optimization problems, the program is structured to replicate cooperated hunting among hawks. HHO has had its share of success in bioinformatics, engineering, and even economics.

After considering both exploratory and exploitative strategies, the HHO algorithm responds in a timely manner and makes a determination for the best solution. Initially, the algorithm takes a random hawks' searching space and starts placing them into it. A solution can be envisioned at each hawk's dispirited position.

In the initial phase of the problem solving, the hawks tend to explore the entire search space as they move around in different directions. This diversity helps prevent the algorithm from converging prematurely. Hawks' movements can also be simulated mathematically using distinct models such as Levy flights for effective exploration of the search space.

The second phase of utilization is called the exploitation phase. Here optimal solutions are located, and subsequently improved according to their fitness values. Position or solutions of Hawks are updated in order to search for better solutions in an iterative system. Various techniques and approaches like operators of crossover and mutation can constrain the updating direction in order to heighten the exploitation of the favorable areas in the search space. Some of the key strategies include:

• Soft besiege- during this time, hawks try to lessen the distance between them and the prey while tightening the area from which the prey can escape.

• Hard besiege- The moment the hawks are closing up to the prey, they all quickly swarm together and capture it.

• Soft besiege with surprise pounce- in the event that the prey has been weakened or cornered hoping they will be the first to strike, the hawks attack at once.

• Hard besiege with surprise pounce- after a rapid pursuit without hesitation pounce on the target closing the distance quickly and capturing it, is the ultimate hawk strategy.

The HHO algorithm imitates the shooting strategy of Harris hawks which is a community approach where everyone works together as a unit. Harris hawks position and share details concerning their locations together with their fitness levels with one another during the cooperative hunting phase. The cooperative and the competition focus on which the algorithm can learn something useful and make decision. Such cooperation mechanism of hunting enables the algorithm to search solution more effectively by making better use of knowledge in the group.

The hawks of the current generation modify their current place where they are located with respect to information acquired during hunting. This provides scope for the evolution of various Adjusters such that the fit function takes higher values than the previous one. The specific guidelines or mathematical functions employed to assist in position updating are related to the problem at hand.

A sequence of steps of searching, seeking target and modifying position is repeated till a condition defined beforehand is fulfilled. A target fitness ratio or a certain number of iterations is a realistic to consider for a stopping criterion, as after that initial progress ceases and only minimal adjustments can be seen to persist. The aforementioned

threshold guarantees that adequate amount of time has been spent exploring the area of concern to come up with an acceptable answer.

Various optimization problems that include function optimization, parameter tuning and subset selection have been tackled by hawk hunting optimization approach. Observing the feeding techniques of hawks means they have learned how to masterfully mix exploration and exploitation so that best solution is found in shortest time possible [23].

2.3.2. Advantages of HHO

1. The Balance between Exploration and Exploitation: As an interesting aspect of HHO, the methodology exhibits a balanced level of exploration and exploitation. A suitable depth of such exploration and exploitation can be observed during the early and the later stages of this algorithm respectively.

2. The Performance at the Convergence Point: It is due to multi-phase techniques and their adaptive hunting strategies that HHO appears to optimally converged faster than some well known optimization algorithms.

3. Effectiveness and Robustness: The HHO mathematical model is relatively straightforward thus its implementation is not a great deal. This makes it readily deployable in a wide range of optimization tasks with very few changes, and thus makes it applicable in a number of disciplines.

4. Locally Optimal Avoidance: The HHO solutions can move through suboptimal zones to the optimal ones by modeling different leaders positioning ways which attack different zones where hounds exist, thus increasing the chances of getting the best solution.

2.3.3. Applications of HHO Algorithm

The Harris Hawks Optimization (HHO) algorithm has been effectively utilized in optimization problems, including bioinformatic areas. This subsection aims to evaluate some of the applications of the HHO algorithm in bioinformatics research.

As already stated, bioinformatics has many vital components and one of them that is significant is the selection of features. This involves selecting the features or variables which are useful for classification or prediction of biological data. HHO has been

employed to conduct feature selection by ranking all the features with respect to their importance or relevance and using the best combination of features which will yield the best maximum classification accuracy or the least prediction errors. The proposed HHO algorithm is shown to be effective for feature selection because of its capability to search for an optimal feature subset [39].

In this context, clustering, which is one of the fundamental bioinformatics tasks pertaining to grouping identical data instances that require bioinformatics, is also of concern. The evaluation of the results in the context of the HHO algorithm can be further enhanced by limiting the interdistance of clusters while widening their internal distance, among other parameters in the process of clustering. The ability to search and to use different regions of the solution space effectively enables the HHO algorithm to identify the best candidates for the cluster centers or assignments, making the clustering results accurate and reliable [23].

Classification, an important task in bioinformatics, is concerned with the assignment of instances to specified classes or categories. The application of HHO has enabled the use of ANN, SVM, Decision trees, and many others as classifiers. The utilization of HHO algorithms has enabled classification tasks to be accomplished more attentive by increasing either the performance or accuracy of the model employed [38].

A number of bioinformatic algorithms and models require the tuning of multiple parameters for enhanced output. HHO has been used to tune parameters in bioinformatic algorithms for instance in DNA sequence alignment, protein structure prediction as well as gene regulatory network inference. By utilizing the HHO algorithm's exploration and exploitation capabilities, parameters can be effectively tuned, hence increasing the accuracy and efficiency of the bioinformed models.

Bioinformatic models such as those that are based on machine learning or statistical models often include parameters which require optimizing in order to generate the required output. HHO has been used to fine-tune the parameters of such models, so as to optimize their performance on specific datasets.

The bioinformatic-based uses of HHO algorithm show its applicability in addressing the optimization issues faced within the bioinformatic field. It is worth noting that the HHO algorithm is capable of complementing the bios tasks as it has an efficient embedded optimization layer due to the innovative hunting techniques of the Harris

hawks. These applications show the effectiveness of the HHO method in solving various optimization problems in bioinformatic. Combining the HHO algorithm with ANN has a great potential for optimizing the processes of classification, clustering, feature selection and parameter optimization and thus improving the practical use of bioinformatic data. Underlying some of the leading causes of mortality across the globe in any given year, heart disease stands out as one of the major contributors. At the beginning of the 20th century, heart disease represented 91% of total mortality on the planet. At the end of the century, the ratio of deaths attributed to cardiovascular diseases reached 52%. Heart disease remains to be the greatest factor contributing towards deaths across the globe. This work initially sought to develop a method for diagnosing cardiac disease through the use of a neural network and Harris Hawks Optimization algorithm [1].

2.4. Hybridization of HHO and ANN

In context, it is necessary to highlight that the objective of unifying Harris Hawks Optimization (HHO) and Artificial Neural Networks (ANN) is to take the advantages of the two approaches in order to enhance the optimization in the relevant area of bioinformatics data processing. Both HHO and ANN approaches are especially useful for modeling complementary research aspects and exhibiting high performance in the search for a solution, since HHO is best suited for managing the scope and exploitation of space, while ANN models the learning and predicting abilities. In this way, researchers will be able to take advantage of HHO in locating the best solution and of ANN in handling bioinformatics complex data modeling. Presently, the aim of the hybrid algorithm is to combine the global search aspect of HHO with the learning and predicting aspects of ANN in order to improve optimization and provide reliable and accurate results in bioinformatics work. The main motivation for the hybridization of HHO and ANN is to alleviate the weaknesses of each method while taking advantage of their synergistic properties in bioinformatics. ANNs can learn to identify and generalize complex patterns across large-scale bioinformatics datasets. However, the performance of ANNs is highly dependent on the propriety of the training data and would easily overfit or get into a local optimum.

Based on their description, HHO appears to be designed as an intelligence taking advantage of its efficient strategy of search to traverse the solution space and scout for the best capabilities in the process. Neural networks, on the other hand, are seen as a

vision model capable of projecting having learned from the patterns in data allowing for better parameter optimization obtained in the process. The combination of HHO and ANN in the field of bioinformatics comes with several advantages. First, it increases the chances of a good balance being struck between exploration and exploitation in the course of the optimization process by integrating the global search capabilities of HHO and the learning capabilities of the ANN. This increases the chances of the algorithm succeeding in finding good quality solutions in a reasonably complex, multiconditioned search areas. Second, the hybrid algorithm integrates the learning and generalization capabilities of ANN so as to improve the prediction and reliability of results in the optimization process. In particular, by learning from the data, the hybrid algorithm is expected to improve significantly its performance in detecting the relevant features of the underlying relationships between variables in order to make accurate judgements.

2.4.1. Previous Studies on Hybridization

The combination of HHO and ANN has been researched by several authors for bioinformatics applications. These works have substantiated the ability of using combinations of these techniques and the results achieved were better in terms of optimization and prediction accuracy. Here we mention two of the latest examples of hybridization of HHO and ANN:

In [40], the authors proposed an approach for gene identification with the use of HHO and ANN for classification of cancers. Informational genes for the classification tasks were identified using cells of the HHO which facilitated the feature selection process. These genes were then used to train the ANN for cancer classification. The proposed hybrid system outperformed the individual methods of classification in terms of accuracy indicating the complementarity of HHO and ANN hybrid model. In [38], the hybrid model first HHO then ANN is presented as a means to enhance the performance of a support vector machine (SVM) classifier applied to gene expression data. The authors have implemented the HHO algorithm in this study to find suitable SVM parameter values, while the ANN was utilized to assess the potential winning solutions. The presented hybrid model yielded better accuracy in classification as compared with the conventional optimization models. These previous studies indicate the capability of the hybrid HHO and ANN for bioinformatics purposes.

The strengths of HHO optimization and ANN learning can be successfully integrated to solve more complex design optimization problems and modeling in the field of bioinformatics. The hybrid algorithm will help to provide better predictions, improved generalization of the model, and enhanced biological insights.

2.4.2. Advantages and Challenges of Hybrid Algorithm

The combination of HHO and ANN is termed the hybrid algorithm which is popularly used in optimization tasks and bioinformatics. Some of the key advantages include:

It is obvious that the hybrid algorithm utilizes both algorithms benefit using HHO of resources as well as ANN. With HHO one can easily hope for developing strong global search and even Heuristic exploitation approaches which help one to search for more resources. ANN works mostly only when there exist policies enabling curves in complex data patterns and then predictions. It stands to algorithm's advantage to merge these and it can be said that developing hybrid algorithms encourages one to improve optimization by endeavors expansion encompassing search space area in them and grasping how to make exhaustive forecasts during various optimization phases.

By capturing generalizable patterns from training data, ANN can also output predictions for new data samples. Adding ANN to the amalgamated algorithm can assist in enhancing the generalization of the optimization processes. Thus, it enables the hybrid algorithm to be able to address the complex structural bioinformatics problems, thus making the algorithm more reliable and efficient as such bioinformatics challenges can be performed on novel available data points.

There are wide ranging biological variations which brings noise, subsequent missing values and sometimes outliers. Noisy data can be circumvented in patterns using ANN algorithms. The inclusion of ANN seems to play an important role in the hybrid algorithm in reducing the negative impact of data anomalies thereby enhancing the optimization of the process. Applications in bioinformatics would gain from this nature of algorithms because of the hussle nature of the data.

Additive in nature, the Incisive HHO is a hybrid algorithm that harnesses the best of both worlds. The incorporation of ANN into HHO significantly increases convergence speed, enhances exploration diversity, and most importantly guarantees convergence global optima. HHO fast convergence also ensures that good solutions are easily

attained. Incorporation of prior aids such as expert knowhow's further guarantees optimum solutions instantaneously. Such unique collaborations will greatly enhance the overall bioinformatics solutions which would otherwise be tedious and traumatizing without expert attachment.

The hybrid algorithm not only provides the freedom of integrating domain knowledge, but also makes it possible to incorporate domain specific constraints into the optimization procedure. Incorporating ANN in such problems enables the hybrid algorithm to make use of bioinformatics knowledge, or any other resource, constraints by means of the network structure or other features. Quite simply, this broadens the scope of applicability of the algorithm to actually real bioinformatics problems and enables one modify the optimization due to the task.

Even though the combined technique does have its pluses, its use and deployment come with certain difficulties. For instance, some of these challenges are:

The combined algorithm fixes the parameters of HHO and ANN components. Though it is worth it, finding an optimal hardware configuration can be quite tricky. It necessitates the extensive tuning of population size, convergence criteria, learning rate, and network structures to measure their suitability. The selection of these parameters can impact the overall performance of the algorithm, and domain knowledge may be required to debug them or, in some instances, large computation time.

The combined approach takes the computational burden posed by HHO and ANN altogether. HHO, on the one hand, entails the calculation of fitness functions for every solution. On the other, the ANN encompasses the training of the network and the modification of its parameters. The combined method is, thus, likely to require more resources than either of its components executed separately. Training an ANN and optimizing an HHO are resource-intensive tasks that consume computer memory and processing power. For most of these combined approaches, it might be preferred to utilize them in instances of low-dimensional or smaller bioinformatics problems due to their high computational costs. The hybrid structure also requires strategic frames to ensure it is easy to use and meets its ideal purpose.

This sort of lack of transparency can prove to be a setback in the understanding of the bio-informatics related problems, considering that it strives to propose a molecular based understanding of biological systems. Solutions can be provided by hybrid

algorithms, but usually one cannot point out the particular type of features or their interactions that lead to derivation of that output. There is a need to start working on how to modify the results of the hybrid algorithm in order to sufficiently account for factors that are relevant and for the hybrid algorithm to be applicable meaningfully.

This mostly increases the difficulty level in the optimization procedure of HHO and ANN models. This is due to the fact that it encompasses communication of two different algorithms that requires the sharing of information, coordination of parameters that leads to communication in the learning and searching activities. In order to correct this, adequate measures need to be taken in order to control the degree of interdependence between the HHO and ANN and the efficient merging of the two as well.

The hybrid algorithm has the potential to be quite effective, however, one central issue arises. That issue may not be the most 'critical' one but several aspects must be considered when applying the Hybrid HHO-ANN. The plethora of issues highlighted could be anything between external market forces to internal idiosyncrasies defining the hybrid algorithm's configuration. When adopting the AI-based optimization within the HHO framework, one recurring issue arises, how best do the HHO and the ANN algorithms interconnect. There is a plethora of concerns that need to be highlighted and this includes selecting appropriate parameters, challenges, convergence, and the overrides to name a few. What these paradoxes represent is nothing but the greater fragmentation in understanding the sequences of outcomes given a set scenario. This fragmentation in understanding the core structure culminates into further fragmentation concerning resource allocation and with it optimization. There does exist interdependent linkage between the parameters cited and feasibility. In other words, when assessing the feasibility and suitability of the hybrid algorithm in targeted bioinformatics cases, there needs to be a balance between cost, speed of convergence, and the quality achieved.

These challenges arise from the If the there is a reasonable optimization around the set objectives resilience of an exemplified algorithm can be implemented. As for targeting optimization around certain bioinformatics related challenges, there need to be considerations of scope and specificity targeting the parameters of focus or the specific algorithms. By removing the rigidity of the parameters mentioned above boundaries can be shifted and, in this case, bioinformatics boundaries can be pushed further.

2.5 Comparing HHO with optimization algorithms

In this section, we will compare useful Jellyfish optimization (JFO) method with the HHO method we have already presented above.

HHO and JFO are two biologically inspired algorithms developed to solve challenging optimization problems. When used in biological data mining, they differ in their performance due to the mechanism involved, particularly in how they handle the complexity of biological datasets that are in most cases high-dimensional, highly noisy, and nonlinear. A better understanding of these algorithms helps in appreciating their performance as well as their potential for other tasks in biological data contexts such as heart disease, liver disease, and iris recognition. The exploration and exploitation strategies are well balanced, making them suitable for complex optimization tasks. Take, for example, the way jellyfish swim using ocean currents. They use two types of locomotion strategies: a passive strategy, which involves the jellyfish simply floating with the ocean currents, and an active strategy, which involves the jellyfish actively swimming toward food. They are largely focused on exploration (i.e., trying to find very large solution spaces), although they can be adapted to optimize the local situation if necessary. It is generally easy to implement, but its exploitability compared to HHO is somewhat low. HHO is known for its high convergence rate and ability to avoid local minima and has been successfully applied in optimizing machine learning model parameters. This often leads to improved diagnostic accuracy and reduced computational complexity, which is good for clinical applications. When it comes to bioinformatics analysis tasks that require both global exploration and local exploitation, HHO appears to be superior to JFO. This is because HHO does not stay in these stages for long as it is relatively dynamic and thus switches between the two modes, making the algorithm particularly suitable for applications such as feature selection, parameter tuning, and other bioinformatics optimization tasks. Although JFO can be used for global exploration analysis, it is recommended that it be hybridized with some other algorithms or require additional modifications in order to exploit in-depth bioinformatics applications. Thus, for precise work such as disease diagnosis or biological network optimization, HHO is more ideal while JFO is ideal for early stages of research.

CHAPTER3: HHO APPLIED FOR HEART DISEASE

Diseases that affect the cardiovascular system (CVD) usually, have the greatest number of fatal cases around the world. As per the statistics presented by the World Health Organization (WHO), around 17.9 million deaths are due to the reason. Effective treatment lets these figures be lowered, and as a result, an early and proper diagnosis is vital for these types of problems. Generally speaking, the field of heart disease heavily relies on scheduled clinicians and multiple laboratory tests, for example, ECGs, echocardiography, or even blood tests. In any case, it can take extra time and multiple tests to validate such methods. Unfortunately, testing isn't always its strong suit and can lead to new technology with the goal of achieving higher results with new forms of diagnostics able to provide much more accurate testing.

This chapter presents the results obtained from integrating the Harris Hawks Optimization (HHO) algorithm with machine learning techniques to diagnose heart disease more effectively, which was published in [1][2][3]. The HHO algorithm is particularly useful for feature selection, which is critical for developing robust machine learning models.

3.1. Introduction

Heart disease is one of the most important diseases that kills many people in the world every year. An erratic heartbeat is known as an arrhythmia. Because there are no symptoms in the early stages of cardiac arrhythmias, the disease is typically not recognized until it has progressed significantly. To treat this condition promptly, an early diagnosis is required. An intelligent system capable of early cardiac arrhythmias detection is necessary. In terms of both severity and frequency of mortality, heart disease ranks first. Still, fewer people will lose their lives to heart disease if doctors catch it early and treat it effectively. One of the ways to diagnose cardiac problems is by analyzing electrocardiogram patterns. The doctor can use these signals as a diagnostic tool since they show the electrical potential the heart generates graphically. Five distinct types of arrhythmias have been detected with the help of artificial networks [41]. CNN convolutional neural networks can be used when information related to heart

problems can be expressed as images. In contrast, Recurrent neural networks (RNNs) can be appropriate when information related to heart problems is processed sequentially.

The HHO algorithm offers several benefits for feature selection in heart disease diagnosis. It balances discovery and exploitation, enabling comprehensive exploration of space. Hawks' hierarchical structure facilitates exploration worldwide while effectively exploiting promising areas. This bypass the drawbacks of traditional techniques, which can be biased or computationally difficult.

In this work, we use a hybrid algorithm that combines HHO and ANN to predict these arrhythmias. We describe two methods: one hybrid method using ANN and the other using CNN.

3.2. Dataset

The Cleveland Heart Disease dataset is frequently utilized for predicting heart disease through supervised machine learning techniques. Sourced from the Kaggle machine learning repository, this dataset was originally compiled in 1988 for a health research study by the Cleveland Clinic. Initially, it included 76 distinct features recorded for 303 individuals. Nevertheless, most researchers typically focus on just 14 of these features, which encompass the target class features. Key metrics in this dataset include age, sex, blood pressure, cholesterol, blood sugar, and various other health indicators. The original dataset categorizes subjects into five class labels, represented by integer values from zero (indicating no disease) to four. Research efforts involving the Cleveland dataset primarily aimed to differentiate between the presence of heart disease (values 1, 2, 3, and 4) and its absence (value 0). Researchers suggest simplifying the five class labels into two categories: 0 for no disease and 1 for disease. The target features denote whether an individual has heart disease [42]. The specific features are as follows:

1) Age: One of the vital features of this dataset is that its value is an integer.

2) Sex: It has two valuesof zero and one, which determine the sex of men and women, respectively.

3) Chest pain type (Cp): chest pain results from transient partial occlusion of the coronary blood flow to the heart muscle. Neo angina reflects impaired myocardial perfusion as flow through the coronary arteries increases demand for oxygen (e.g.

during stress or exercise). If chest pain either at rest or provoked by exercise does not resolve within several minutes, the likelihood of acute myocardial infarction is increased significantly, and the patient should proceed to the nearest medical facility. Angina is usually temporary and sometimes chronic, persistent pain. In this dataset for the angina field, the values of Typical angina, typical angina, Non-anginal pain, and Asymptomatic angina are considered, which are indicated by the numbers 1, 2, 3, and 4, respectively.

4) Resting blood pressure (Trestbps): This feature measures a person's blood pressure at rest, usually in a hospital. Before measuring this feature, the doctor ensures the person is not physically active.

5) Serum cholesterol (chol): Serum cholesterol measures the amount of cholesterol in the blood, including HDL, LDL, and some other blood fats. In a healthy person without other cardiovascular risk factors, serum cholesterol is less than 200 mg/dL. This property is continuous and numerical and is expressed in milligrams per milliliter.

6) Fbs: Fasting blood sugar is measured in milligrams per milliliter. The value will be one if it is more than 120 mg/ml. Otherwise, it will be zero.

7) Resting electrocardiographic results (Restecg): This feature has values of 0, 1, and 2 depending on the shape and curve of the electrocardiographic chart.

8) Thalach: Maximum heart rate achieved.

9) Exercise-induced angina (Exang): Angina pectoris is a condition in which a patient develops chest pain that is the source of pain in the coronary arteries. Angina pectoris is caused by a lack of oxygen in the heart muscle. This pain is mainly in the middle of the left chest and can spread to the left arm. It sometimes has two arms and components, such as the jaw and the middle part of the two shoulders. Exercise-induced angina pectoris occurs when a patient develops this pain based on relatively intense exercise. Depending on the presence or absence of this pain, this characteristic has two values : one and zero.

10) Oldpeak: Stress test depression induced by exercise relative to rest.

11) Slope: Reduced ST time during activity compared to resting state. With values of 1, 2, and 3, we can see that the peak workout ST segment has three different ways of sloping: flat, mild slope, and steep.

12) Major vessels (Ca): This feature is possible when using color images and can range from zero to three.

13) Thallium heart rate (Thal).

3.3. HHO Algorithm Based on Artificial Neural Network for Heart Disease Diagnosis

3.3.1. Methods

This work uses the HHO optimization method to select features based on the neural network with the highest accuracy to predict heart disease. Several techniques, such as Multilayer Perceptron (MLP), Support Vector Machine (SVM), Random Forest (RF), AdaBoost, and Bayesian Network, have been implemented and their performance compared. Each feature's importance score is analyzed and estimated using all the methods mentioned. All the features are ranked based on the importance score to find the highest predictivity features.

3.3.1.1. The steps of the proposed method for diagnosing heart disease are outlined below.

a) Pre-processing

A vital part of any healthcare system is the data collected from patients. The data must have a suitable structure and format for machine learning. The following phases and steps are performed in the preprocessing of data related to the treatment system:

- Data that is empty or has an empty value was ignored or filled based on the average value of that attribute or field.
- All data convert to numerical data.
- The collected data is normalized to be ready for machine learning.
- The data used in machine learning and data mining has a set of features with a specific range.

Some can change at a small interval and some at a considerable interval. Using features with a different range of variations can reduce the learning accuracy, so normalization is used in this study to normalize medical data analysis. To normalize patient-related data, the normalization range [a, b] can be considered, and the data can be normalized according to (1):

$$normal(F_i) = a + \frac{F_i - min}{max - min}(b - a) \quad (1)$$

in this equation, F_i is the non-normalized value of a feature, and $normal(F_i)$ is the normalized value of the feature F_i. The max and min values are the maximum and minimum values of the features of a column of the dataset, respectively.

If the normalization interval is considered equal to [0,1], then normalization is performed as (2):

$$normal(F_i) = \frac{F_i - min}{max - min}. \quad (2)$$

In the proposed method, a binary feature vector X_i with n components such as (3) define a member i of the HHO algorithm. The value of each component X_i^j is zero or one, which indicates the lack of feature selection and feature selection, respectively.

$$X_i = \ll X_i^1, X_i^2, X_i^3, \ldots, X_i^n \gg. \quad (3)$$

A feature vector in the current iteration is $X(t)$. This feature vector is $X(t + 1)$ in the next iteration. To evaluate any feature vector, it can be mapped to the data and considered the input of an artificial neural network (classifier). Each feature vector is assessed based on the network diagnostic error and the number of selected features for all features. The appropriate objective function for feature selection is defined as:

$$f = \alpha * \frac{1}{n}\sum_{i=1}^{n}|\bar{Y}_i - Y_i| + \beta * \frac{F}{A}. \quad (4)$$

In the objective function, the distance between the actual value, Y_i, and the predicted value, $\bar{Y}_i$, is promediated for the n samples. F and A values are the number of selected features and the total number of possible features, respectively. The coefficients α and β are two random numbers with values between zero and one, and their sum equals one.

b) Feature selection based on HHO

The HHO algorithm is a meta-heuristic algorithm with a swarm intelligence approach and modeling on the behavior of Harris Hawks in nature. This algorithm has swarm intelligence behaviors with a hunting approach. Harris Hawks can detect various chase patterns based on the dynamic nature of hunting scenarios and escape patterns. The Harris Hawks' hunting behavior is shown in Figure 3.1 [43].

Figure 3.1: Mechanism of swarm intelligence hunting in Harris Hawks Optimization (HHO).

The problem has a solution in the form of prey, and that solution is Harris hawks. As things stand, the rabbit posture is the best option. The HHO algorithm incorporates a phenomenon known as silent or gentle siege. Harris hawks will approach their prey cautiously and quietly while searching all around it, a technique known as a silent siege. In the heat of battle, a Harris hawk might dive headfirst toward the rabbit. Figure 3.2 displays the results of the hard siege modeling [43].

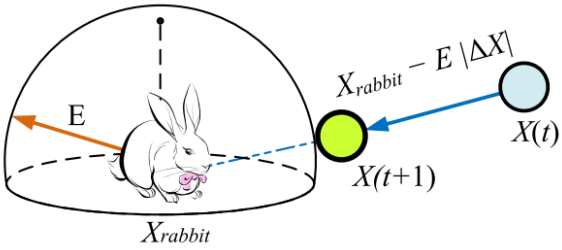


Figure 3.2: Hard siege behavior in the HHO algorithm [43].

During a soft siege, Harris hawks swiftly circle around their target. In this strategy, the hawks approach the prey when the timing is right. Their behavior includes diving from a height and then flying away from the target. As they descend, they gradually move closer to the prey. In HHO algorithms, each hawk can adjust its flight path based on the gathering center of other Harris hawks. Figure 3.3 illustrates the preferred behavior for a Harris hawk moving towards a mean point. By calculating the average position of the

population alongside their optimal positions, the Harris hawk gently approaches the prey during the siege [43].

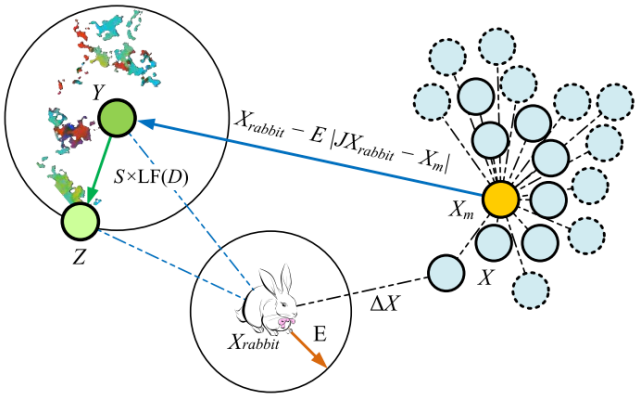


Figure 3.3: Rapid dive behavior in the HHO algorithm [43].

The position of the Harris hawks and the rabbit, or the current optimal solution, is constantly updated by iteration of the HHO algorithm. In the last iteration, the prey position is the optimal solution. In each iteration, this algorithm attempts to update the feature vectors. The HHO algorithm tries to minimize the value of the objective function *f*, updating its value in each iteration.

The optimal feature vector is displayed in each iteration with $X_{rabbit}(t)$. (5) is used to update feature vectors with random motions:

$$X(t+1) = \begin{cases} X_{rand}(t) - r_1 |X_{rand}(t) - 2r_2 * X(t)|, & rand \geq 0.5, \\ (X_{rabbit}(t) - X_M(t)) - r_3(LB + r_4(UB - LB)), & rand < 0.5. \end{cases} \quad (5)$$

The value of $X_{rand}(t)$ is a random position of a feature vector in the problem space. The value of $X_M(t)$ is the point of gravity and the characteristic vectors, r_1, r_2, r_3 , and r_4 are uniform random numbers in the range of zero to one. The *LB* and *UB* parameters are the lower and upper ranges of solutions in the problem space, respectively. The values of the *LB* and *UB* parameters of the proposed method are zero and one, respectively, and therefore (5) becomes (6):

$$X(t+1) = \begin{cases} X_{rand}(t) - r_1 |X_{rand}(t) - 2r_2 * X(t)| & rand \geq 0.5, \\ X_{rabbit}(t) - X_M(t) - r_3 * r_4 & rand < 0.5. \end{cases} \quad (6)$$

By updating the feature vectors under the search agent, the feature vectors update in subsequent iterations under the influence of another type of search called a soft besiege, which is shown in (7):

$$X(t+1) = \left(X_{rabbit}(t) - X(t)\right) - E|J * X_{rabbit}(t) - X(t)|. \quad (7)$$

In this equation, *J* is a random value between zero and two. The coefficient *E* is also a parameter called the energy coefficient and is a decreasing factor in terms of iteration. Another type of update is related to modeling Harris hawk dives and can be used to update feature vectors, the modeling of which is shown in (8):

$$X(t+1) = X_{rabbit}(t) - E|X_{rabbit}(t) - X(t)|. \quad (8)$$

In HHO algorithms, each feature vector can be updated based on the average population position or population center of gravity, as in (9):

$$X(t+1) = X_{rabbit}(t) - E|J * X_{rabbit}(t) - X_m(t)|. \quad (9)$$

By applying these relationships, feature vectors are updated in each iteration to diagnose the disease. The most optimal feature vector reduces the disease's diagnostic error in the last iteration.

3.3.2. Results and discussion

This section analyzes the proposed method for diagnosing heart disease. The learning methods are evaluated using MATLAB software version 2019b.

3.3.2.1. Evaluation criteria

To evaluate the proposed method, the parameters of true positive, true negative, false positive, and false negative are used, which are explained below:

- *TP*: Subjects with atherosclerosis and correctly predicted by the proposed method as "heart patients".
- *TN*: subjects without atherosclerosis and correctly predicted by the proposed method as "healthy".
- *FP*: Subjects without atherosclerosis and wrongly predicted by the proposed method as "heart patients".
- *FN*: subjects with atherosclerosis and wrongly predicted by the proposed method as "healthy".

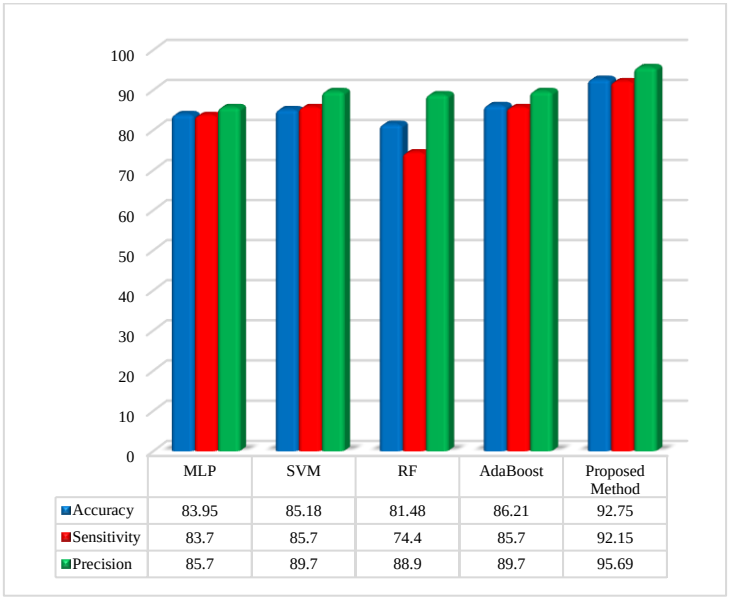
Accuracy, sensitivity, and precision, defined as follows, are used to assess the model [44][45][46][47].

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (10)$$

$$Sensitivity = \frac{TP}{TP+FN} \times 100\% \quad (11)$$

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (12)$$

Figure 3.4 compares accuracy, sensitivity, and precision indicators for diagnosing heart disease with the proposed method and other methods.



	MLP	SVM	RF	AdaBoost	Proposed Method
Accuracy	83.95	85.18	81.48	86.21	92.75
Sensitivity	83.7	85.7	74.4	85.7	92.15
Precision	85.7	89.7	88.9	89.7	95.69

Figure 3.4: Comparison of accuracy, sensitivity, and precision of the proposed method and other methods in diagnosing heart disease.

3.3.2.2. Analysis and classification of heart risk

In this section we evaluate the method in evaluating the heart risk. The output feature of this Cleveland dataset has five different classes that indicate the possibility of clogged arteries in the heart vessels. with 0 corresponding to no heart disease, and ranges from 1 to 4 from least to greatest severity (1: Mild or minimal heart disease, 2: Moderate heart disease, 3: Severe heart disease, 4: Very severe heart disease). We reduce the five-class features of this dataset to two classes.; 0 = no disease and 1 = disease. The target feature refers to the presence of heart disease in the subject. Table 3.1 shows the features included in the Cleveland heart disease dataset.

In the original dataset, there are a total of 6 samples with null values; 4 samples in the "Ca (number of major vessels)" feature and 2 samples in the "Thal (thallium heart rate)"

feature. Since there are very few null values, these samples can be removed from the dataset.

Table 3.1: List of features in the Cleveland heart disease dataset.

Order	Feature	Description	Feature Value Range
1	Age	Age in years	28 to 76
2	Sex	Gender	Value 1 = male Value 0 = female
3	Cp	Chest pain type	Value 0: typical angina Value 1: atypical angina Value 2: non-anginal pain Value 3: asymptomatic
4	Trestbps	Resting blood pressure (in mm Hg on admission to the hospital)	94 to 190
5	Chol	Serum cholesterol in mg/dL	125 to 563
6	Fbs	Fasting blood sugar > 120 mg/dL	Value 1 = true Value 0 = false
7	Restecg	Resting electrocardiographic results	Value 0: Normal Value 1: having ST-T wave abnormality Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8	Thalach	Maximum heart rate achieved	74 to 220
9	Exang	Exercise-induced angina	Value 1 = yes Value 0 = no
10	Oldpeak	Stress test depression induced by exercise relative to rest	0 to 6.2
11	Slope	The slope of the peak exercise ST segment	Value 0: upsloping Value 1: flat Value 2: downsloping
12	Ca	Number of major vessels	Number of major vessels (0–3) colored by fluoroscopy
13	Thal	Thallium heart rate	Value 0 = normal; Value 1 = fixed defect; Value 2 = reversible defect
14	Target	Diagnosis of heart disease	Value 0 = no disease Value 1 = disease

In the assessment of the methods, 70% of the dataset is allocated for training, while 30% is designated for evaluation. During the feature selection stage, a population of 20 falcons or feature vectors is utilized, with 50 iterations conducted, and the experiments

are repeated and averaged over 30 trials. The evaluation phase of the feature vectors employs a two-layer artificial neural network, where each layer consists of 20 artificial neurons. To assess the proposed method, the RMSE and MAE error indices are applied, as defined in equations (13) and (14).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{13}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{14}$$

in these relationships, y_i is the actual class number of a person in terms of illness or health, and $\hat{y}_i$ is the class number estimated by the proposed method for that person's condition and *n* is the number of samples used in the evaluation. Figure 3.5 compares the RMSE and MAE indexes for the proposed and other methods.

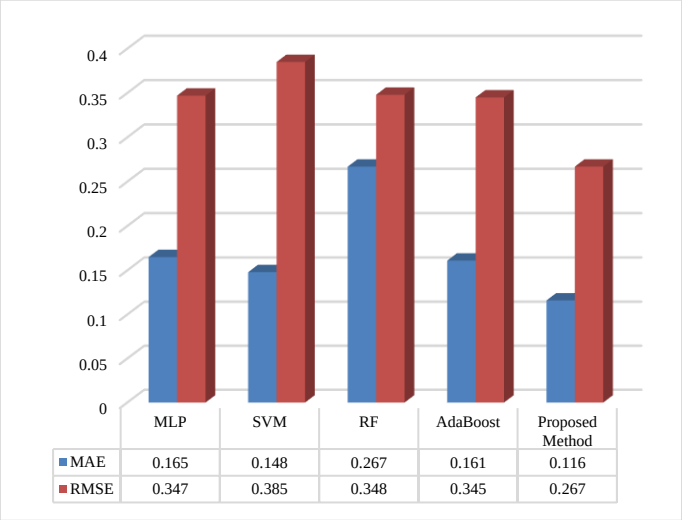


Figure 3.5: Comparison of RMSE and MAE error of the proposed method and other methods.

Analysis experiments show MAE error of artificial neural network, support vector machine, decision tree, random forest, AdaBoost, Bayesian network, and proposed voting by HHO algorithms equal to 0.165, 0.148, 0.267, 0.161 and 0.116, respectively. The proposed method has the lowest error regarding the MAE index for diagnosing heart disease by feature selection. The techniques discussed in diagnosing heart disease with RMSE error are 0.347, 0.385, 0.348, 0.345 and 0.267, respectively. The proposed method with the feature selection mechanism of the compared methods has less error in diagnosing heart disease.

3.4. Patient Privacy in Smart Cities with Blockchain Technology and Block Feature Analysis with HHO Algorithm and Machine Learning

In the future's smart cities, medical centers require confidentiality and data security to provide accurate patient care. The most straightforward approach for sharing medical data involves transmitting information to other facilities without maintaining confidentiality, which is inadequate since safeguarding medical privacy is a fundamental principle in healthcare. The suggested approach consists of two layers. The first layer involves employing blockchain technology for data transmission. The second layer entails analyzing blocks associated with patient records through various methods. Patient records are organized into a blockchain block and sent to other medical centers. Each treatment facility can propose a specific treatment type along with the blockchain attachment, disseminating this information to all nodes and treatment centers. Medical centers then receive patient data and utilize data mining techniques to forecast treatment options. The implementation of this proposed system demonstrates that blockchain technology maintains data confidentiality at nearly 100%. Additionally, the reliability of this framework significantly surpasses that of centralized methods.

3.4.1. Review

Blockchain technology was first developed to safeguard digital currency known as bitcoins. Now, blockchain has matured into a foundational technology for numerous decentralized industries. Information security, healthcare, and insurance are three areas that can benefit greatly from blockchain technology's ability to manage sensitive data. There are four basic parts to health care: study, insurance, treatment, and medical and health service users, sometimes known as patients. Data breaches involving patient information are becoming more common in the healthcare industry annually. An estimated 37 million patient records were compromised between 2010 and 2017, with over 300 violations detected in 2017. Concerns regarding the secure custody, transfer, and use of patient records and medical data have been validated by the healthcare industry's transition to digital records [48]. The blockchain provides a solution to the key issues confronting healthcare. It addresses the secure sharing of medical records and adherence to data privacy regulations. The healthcare industry is continuously

evolving. In health applications, blockchain technology is utilized to ensure privacy. A review of several studies has been conducted that focus on this privacy approach using blockchain. [49] discusses the integration of the Internet of Things and blockchain in health initiatives, examining various healthcare programs that incorporate both technologies. This analysis looked into six medical service programs, which include remote patient monitoring, management of electronic medical records, disease prediction, patient tracking, drug tracking, and efforts to combat infectious diseases, particularly COVID-19. Additionally, the study explores the challenges related to implementing blockchain technology within IoT-based systems and presents some potential solutions. Recommendations for future research have been suggested that could transform the healthcare sector by leveraging other technologies such as artificial intelligence, big data, fog computing, and cloud computing. The potential challenges of the future, including scalability, storage capacity, blockchain size, global interoperability, and standardization, are addressed in [50]. This study emphasizes viewpoints related to health data, the sharing process, clinical trials, the pharmaceutical sector, big data, artificial intelligence, as well as issues of security and privacy. To gauge the healthcare industry's preparedness to embrace blockchain technology, [51] presents a framework. They laid forth a system for gauging the health industry's preparedness to use blockchain technology. All key players are a part of their framework, which accounts for the intricate interplay of different elements, social systems, and institutional mechanisms. The healthcare sector in the UAE can benefit from their suggested structure. The findings highlight the various ways in which government preparedness is crucial to the blockchain endeavor. Evidence suggests that bigger corporations are more open to embracing blockchain's benefits. Everyone involved isn't prepared since blockchain legislation, privacy concerns, and trust issues aren't well-defined. A healthcare system that uses blockchain technology to diagnose diabetes was introduced in [51]. An increasing number of people are losing their lives to diabetes, a chronic disease that is spreading at a rapid pace. This research lays the groundwork for using blockchain cryptography to make diabetes diagnoses. The suggested approach for early illness diagnosis makes use of many categorization algorithms. An electronic health control system that securely records patient information is the proposed framework. Their shared framework combines symptom-based disease prediction, blockchain, and interdepartmental file systems. In this context, patient health information is collected through wearable sensors. The collected

data of the patient is eventually sent to the proposed system administrator to implement a machine-learning model for further processing. The results and the physiological parameters are stored in the data blocks of the blockchain with the approval of the patients and their physicians. In [51], securing blockchain-equipped systems in health care through a deep belief network has been introduced. Cybersecurity is a vital issue and challenge in healthcare due to the legal and ethical perspective of patient medical data. Designing a data confidentiality model for healthcare applications requires special attention to ensure data security. This work proposes a secure intrusion, blockchain-based data transfer identification with a classification model to maintain confidentiality in the healthcare sector. The proposed model uses sensor devices to perform the information acquisition process. The diagnosis was made using the deep belief network model in this context. Blockchain technology securely transfers data to the cloud server, and the ResNet classification model is implemented to detect disease on cloud servers.

There was an investigation on blockchain technology's function in telehealth and telemedicine in [52]. Diseases like COVID-19 can be better controlled with the help of telemedicine and other forms of remote medical care. In hospitals, these health practices can help manage limited health resources and minimize the number of COVID-19 patients. According to their findings, blockchain technology enables secure, decentralized, transparent, traceable, trustworthy, and anti-manipulation telemedicine and remote health services. These advancements in technology have made it possible for medical experts to identify fake medical records and home diagnostic test kits.

In [53], a method for integrating blockchain technology and a dependable information management system for veterinary clinics utilizing predictive analysis were introduced. The healthcare sector is experiencing a significant transformation due to recent developments in information management systems alongside machine learning algorithms. Nonetheless, these systems face multiple issues, including security, reliability, and ease of use. There is a need for a novel solution to enhance data accessibility, and it is essential to appropriately modify security policies. This study intends to employ machine learning algorithms for prediction and authentication through blockchain technology. A health program utilizing blockchain technology for diabetes prediction was proposed by [54], incorporating fog computing methods. This approach gathers patient health data from fog nodes and records it on a blockchain. Initially, a new rule-based clustering algorithm is employed to group patient health

records. Subsequently, an adaptive feature-based fuzzy neural inference system is used to diagnose diabetes and heart disease. The results from experiments indicate that this proposed method effectively identifies the disease, demonstrating an accuracy of approximately 81%, which surpasses that of the neural network algorithm. In [55], the authors proposed a safe model for transmitting and detecting medical images with the help of blockchain on the Internet. This research presents in-depth learning using blockchain image transfer and a secure detection model for the Internet of Things environment. The proposed model includes several processes: data collection, secure exchange, hash value encryption, and classification. First, Elliptic curve cryptography (ECC) is applied, and the optimal ECC switch is generated using the hybridization of the grasshopper with the fruit fly optimization (GO-FFO) algorithm. Their method uses a deep belief network to classify and diagnose diseases. [56] presents an intelligent IoT-based healthcare framework using blockchain technology with an optimal deep-learning-based secure blockchain (ODLSB) model. This work provides a secure blockchain model based on deep learning. The proposed model includes three main processes: secure trading, hash value encryption, and medical diagnosis. Their proposed method consists of an Orthogonal particle swarm optimization (OPSO) algorithm for secretly sharing medical images, the hash value encryption process uses the Neighborhood indexing sequence (NIS) algorithm and uses the optimal deep neural network (ODNN) as a classification model to diagnose diseases. In [57], blockchain-based image steganography and the PSO algorithm are introduced. In this work, modeling is a new way to ensure the updating and sharing COVID-19 data in decentralized hospitals. Updating and securing the sharing of large amounts of healthcare information between hospitals is challenging. There are two issues related to the confidentiality and integrity of health data. Network security vulnerabilities may be a concern for data availability. According to the authors, no study provides safe updating and sharing solutions for large amounts of healthcare information in hospital communication channels. Therefore, this work proposes and discusses a new method based on steganography as a solution. The first step estimates each image's embedding capacity before hiding. The second step is to conceal the COVID-19 data using the PSO algorithm. The third step is image transfer based on blockchain technology.

A review of studies shows that most recent research has used the blockchain to transmit medical data. Studies have used the blockchain to store patient data and records in a

distributed network. One of the challenges of studies in this area is that blockchain nodes do not process patient data to share. In other words, in the reviewed studies, if a patient file needs to be analyzed by medical centers, then these hospital centers need to share their analysis with other nodes. Our contribution to this research is to present a data analysis approach based on machine learning, feature selection, and blockchain technology. The proposed method innovation is to diagnose a person as ill or healthy based on the analysis of all nodes participating in the blockchain. The blockchain needs the other nodes' approval if a node wants to analyze and comment on a patient file. In the continuation of this section, the proposed method and structure used in it and its phases, such as data storage in the blockchain, feature selection, and learning based on the majority vote, will be described.

3.4.2. The proposed method

In most cases, it is necessary to use the advice of several medical centers or hospitals to treat a person. One way to improve the treatment process for patients is to share patients' records with other treatment centers. Each medical center should apply its analysis to the data and then provide analysis to different medical centers. A medical center can use patients' opinions to determine the optimal treatment for them and most of the types of treatment. In the proposed method, each treatment center uses a learning algorithm to diagnose the type of disease. Machine learning algorithms such as decision trees, random forests, support vector machines, multilayer neural networks, AdaBoost, and Bayesian networks are used for analysis in medical centers. The machine learning algorithm in each medical center plays a role in stimulating the opinions of the doctors of that center in diagnosing the disease. These opinions and consultations are shared in medical centers. The recommended method for each hospital is to use blockchain to store and send information and increase data confidentiality. The proposed framework for diagnosing the disease and maintaining data confidentiality has several steps as follows:

- Storage and transmission of patient data by blockchain.
- Analyze blocks of information of patients with chain blocks by machine learning and feature selection.
- The next section will explain.
- Diagnosis of the disease in a medical center by a majority vote.

In the proposed framework, heart disease is used in the analysis. Patients' records contain information related to heart patients, and this data is used for analysis.

3.4.3. Framework of the proposed method

The framework of the proposed method for diagnosing heart disease and sending data confidentially with blockchain is shown in Figure 3.6. Its purpose is to send patients' medical information and records to other medical centers in the smart city. With this mechanism, the opinion of doctors in different medical centers is well received and consider the best treatment. An excellent way to do this is to use blockchain encryption technology. In the proposed method, blockchain technology is used to send patients' files and confirm the type of treatment. In the proposed method, each medical center uses a data mining method to predict the appropriate treatment by receiving information and patient records. In the proposed method, data mining techniques play the role of information analysis in treatment centers. In the proposed method to discover the best treatment for patients, according to Figure 3.6, the following steps are performed:

- Patient information and their records are collected.
- Patient records are preprocessed; normalization is critical in data preprocessing.
- The data is sent as a block of blockchain technology to medical centers in intelligent cities, and each medical center receives it.
- Data of each block received by a healthcare system. Data analysis by machine learning and data mining to predict disease progression.
- The patterns discovered by each treatment center are added to the block by machine learning and sent to other treatment centers by the blockchain.
- To increase the accuracy of machine learning methods in each medical center, the feature selection phase is used in each medical center.
- The HHO algorithm is optimized to select a feature from. This algorithm is used in many applications, and its accuracy in finding the optimal solution is remarkable. The role of the HHO algorithm is to optimize the selection of essential features of heart patients.
- Each medical center confirms and authenticates a block and sends it to the primary treatment center

- The primary treatment center can receive the recommendations of the treatment centers by receiving the block and uses the majority voting mechanism to decide on the type of treatment.
- Each treatment center can use a machine learning technique; ultimately, the majority vote is used. In the proposed method, each medical center can apply the doctors' opinion about the data or analyze it by learning their machine by receiving the data of each block.

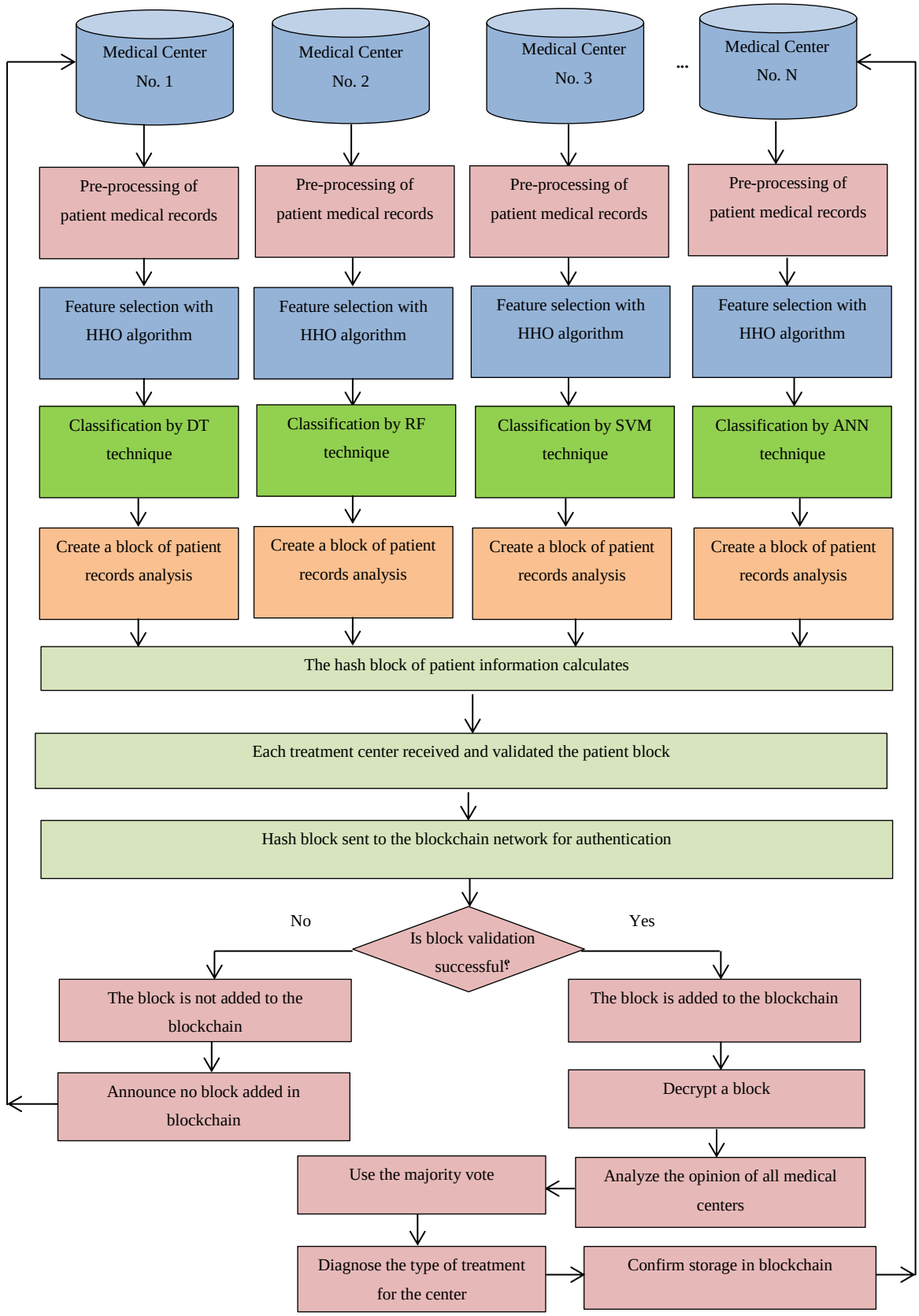


Figure 3.6: Framework of the proposed method for diagnosing heart disease and maintaining the confidentiality of patients' records.

3.4.4. Steps of the proposed method

The steps of the proposed method for diagnosing heart disease are described below. The structure of the blocks used in the blockchain describes preprocessing, feature selection and majority-based learning.

1. The structure of each block

Patient information must be specified in a specific format for submission in the blockchain. According to Figure 3.7, each block contains at least information such as the unique number of each patient, the exceptional number of each hospital, patient information and data, the time of the creation of the block, the hash of details of each block, and the order of placement in each block.

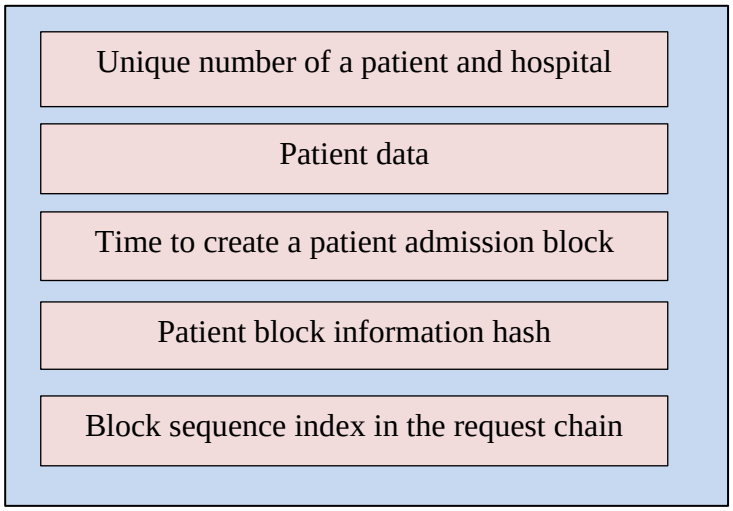


Figure 3.7: Coding of patient information in a block.

The following components are used in each block belonging to the blockchain:

- Block number must be unique.
- The label indicates the date the block was created and can be made with the TimeStamp operation.
- The hash encrypts a block's contents until each bit is manipulated; the hash information is changed, and the block is removed from the blockchain.
- A hash index that specifies which blocks a chain connects to.
- Data related to patient record information, part of which is predicting patients' status through data mining methods.

Blockchain technology is a distributed cryptographic method. Each block is sent to other members after being created as a blockchain member. Each medical center analyzes medical data using its data mining method, sends the final analysis to all groups in the blockchain, and consults their medical opinion with other nodes. The proposed SHA-256 notification method uses the SHA-2 cryptographic hash function family. A collection of heart patient data UCI is used for patient-related information. Its features (the features that are placed in the blockchain data block) are as follows:

1) Age, 2) Sex, 3) Cp, 4) Trestbps, 5) chol, 6) Fbs, 7) Restecg, 8) Thalach, 9 Exang, 10) Oldpeak, 11) Slope, 12) Ca and 13) Thal. This dataset follows a similar structure to that described in Section 3.2.

By applying the relationships between function (1)-(9), feature vectors are updated in each iteration to diagnose the disease. The most optimal feature vector reduces the disease's diagnostic error in the last iteration. In the proposed method, each Harris hawk is a feature vector and contains components zero and one, which indicate the lack of feature selection and feature selection, respectively. On the other hand, the rabbit refers to the optimal feature vector. The objective function evaluates each feature vector with the error of disease diagnosis and number of features.

Figure 3.8 shows the flowchart of the feature selection process for diagnosing heart disease.

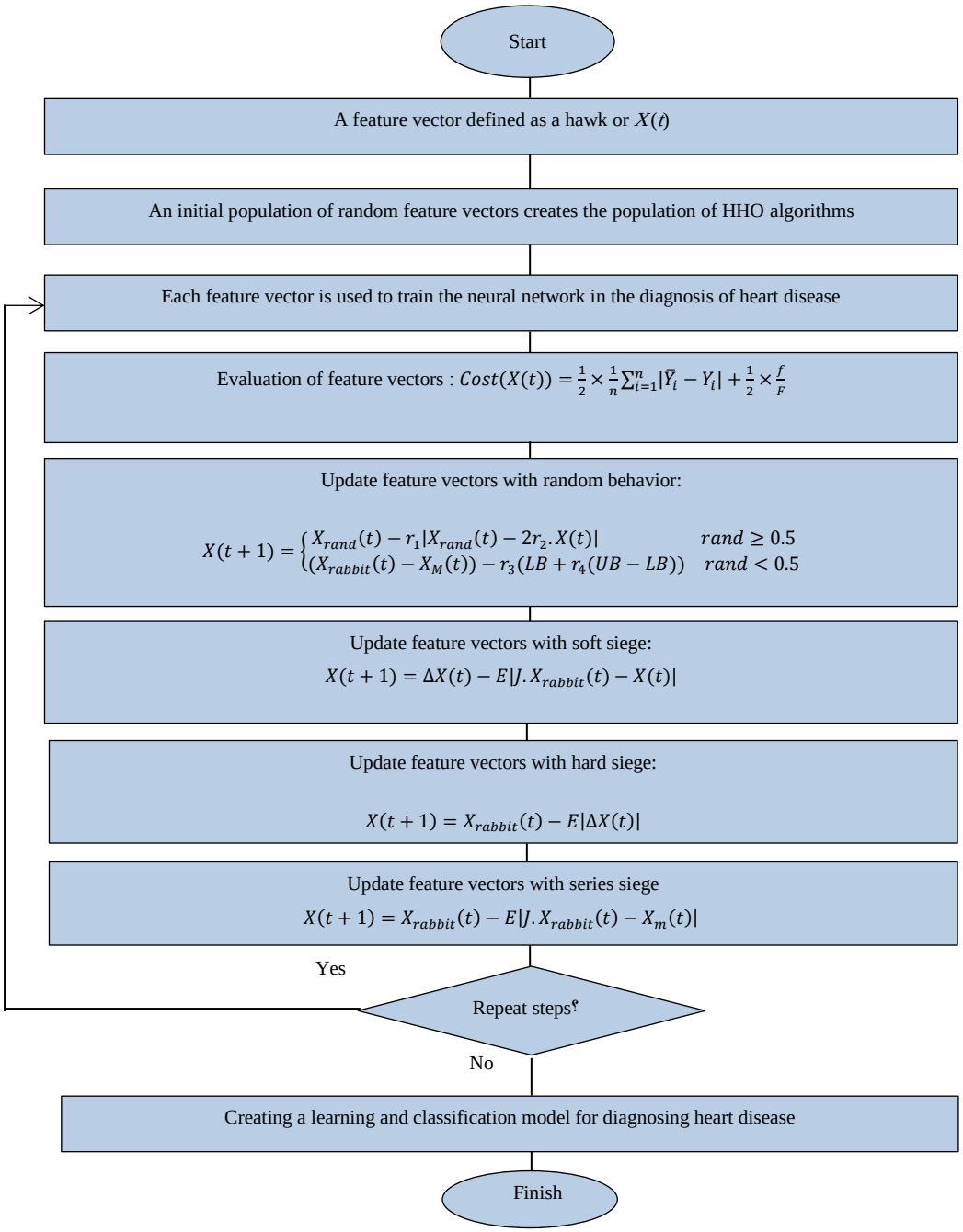


Figure 3.8: Feature selection in blockchain-related blocks.

2. Majority voting

Figure 3.9 shows the framework for using voting-based learning in the proposed method in each blockchain:

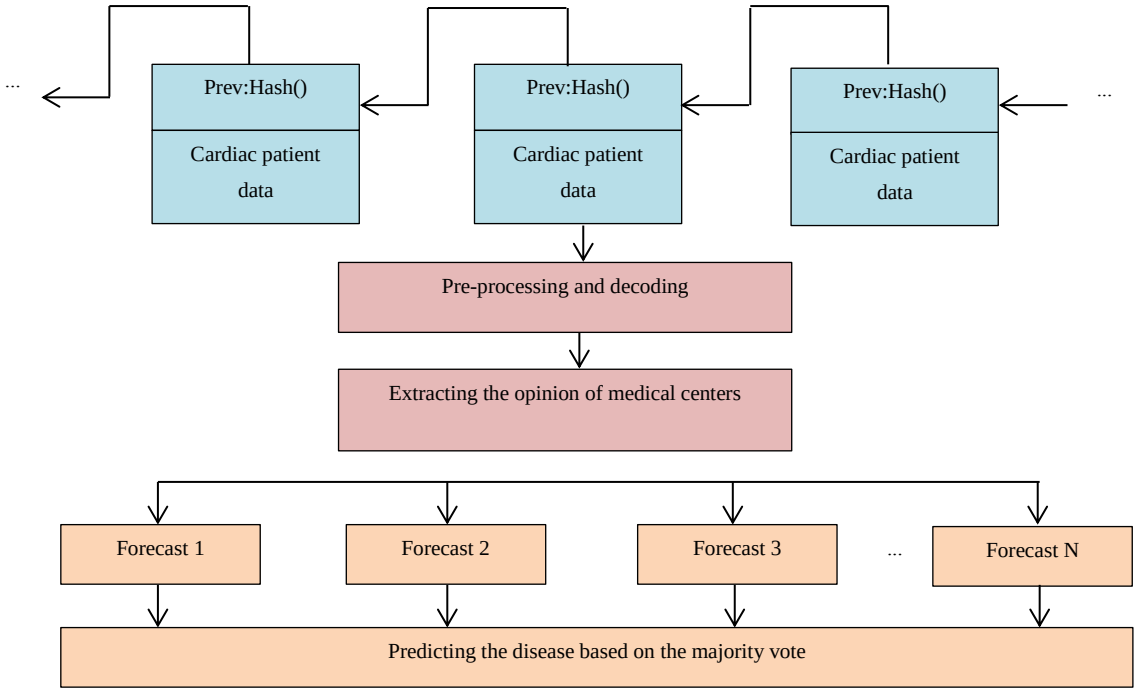


Figure 3.9: Block information extraction and prediction based on the majority vote.

Each hospital and treatment center should make the final analysis of the data based on the majority vote. If a class number is generated in the output of most learning methods, it is selected as the final output. This work uses several methods for majority voting, such as artificial neural networks, support vector machines, decision trees, random forests, Bayesian networks, and AdaBoost. A patient-related block is first selected, and information, such as patient records and other medical centers' opinions about the patient, is extracted. The contents of the data block are extracted from the blockchain. At this stage, the required preprocessing is performed on the block. Disease-related prediction class number (medical centers add this prediction number) is determined by a majority vote.

The type of artificial neural network in the evaluation phase of feature vectors is also selected from the two-layer type, and each layer has 20 artificial neurons. One method to evaluate the proposed method is using the RMSE and MAE error indices, formulated according to (13) and (14). In these relationships, y_iIs the actual class number of a person in terms of illness or health, and $\hat{y}_i$He class number estimated by the proposed method for that person's condition can be healthy or non-healthy., On the other hand, *n* is the number of samples used in the evaluation. Table 3.2 shows the

methods' RMSE and MAE index values and the proposed method for diagnosing heart disease. Figure 3.10 shows the mean error of RMSE and MAE of each of the methods in diagnosing heart disease:

Table 3.2: Comparison of RMSE and MAE error of the proposed method and other methods

Method	MAE	RMSE
MLP	0.165	0.347
SVM	0.148	0.385
J48	0.253	0.427
RF	0.267	0.348
AdaBoost	0.161	0.345
BN	0.169	0.341
Majority Vote	0.128	0.294
Majority VoteHHO	0.116	0.267

Analysis experiments show MAE error of artificial neural network, support vector machine, decision tree, random forest, AdaBoost, Bayesian network, proposed voting, and proposed voting by HHO algorithms equal to 0.165, 0.148, 0.253, 0.267, 0.161, 0.169, 0.128 and 0.116, respectively. The proposed method has the lowest error regarding the MAE index for diagnosing heart disease by feature selection. The techniques discussed in diagnosing heart disease with RMSE error are 0.347, 0.385, 0.427, 0.348, 0.345, 0.341, 0.294, and 0.267, respectively. The proposed method with the feature selection mechanism of the compared methods has less error in diagnosing heart disease.

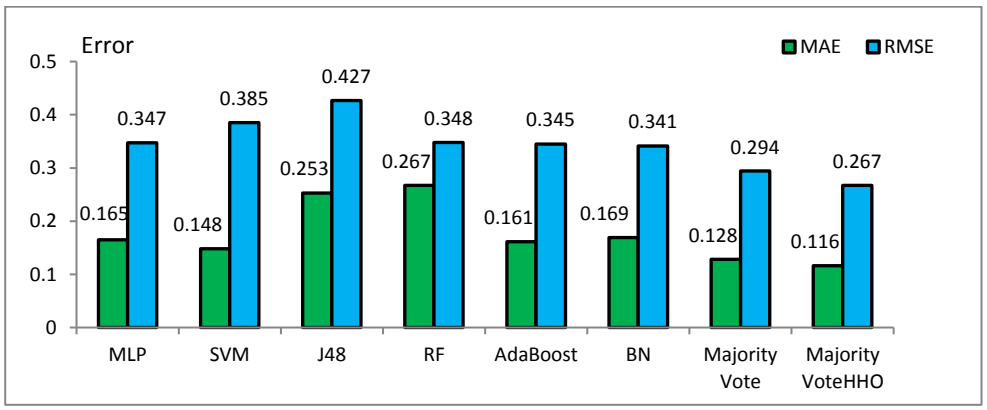


Figure 3.10: Comparison of RMSE error in the diagnosis of heart disease.

Among the learning methods without a voting mechanism, the support vector machine method has the lowest error in the MAE index. In the RMSE index, Bayesian network

error also performs better than non-voting methods in diagnosing heart disease. Classification indicators such as accuracy are of particular importance in addition to error indicators in diagnosing heart disease and analyzing data blocks in the blockchains.

Table 3.3 compares three indicators of accuracy, sensitivity, and precision for diagnosing heart disease in the proposed method and other methods. The accuracy, sensitivity, and precision index of the proposed method and other methods are compared in the diagrams of Figures 3.11, 3.12, and 3.13.

Table 3.3: Comparison of accuracy, sensitivity, and precision of the proposed method and other methods in diagnosing heart disease

Method	Accuracy	Sensitivity	Precision
MLP	83.95	83.7	85.7
SVM	85.18	85.7	89.7
J48	77.78	62.8	93.1
RF	81.48	74.4	88.9
AdaBoost	86.21	85.7	89.7
BN	86.42	83.7	90
Majority Vote	91.87	91.67	93.14
Majority VoteHHO	92.75	92.15	95.69

The experiments' analysis showed that the accuracy of the artificial neural network, support vector machine, decision tree, random forest, AdaBoost, Bayesian network, proposed voting and proposed voting by feature selection in diagnosing heart disease are 83.95%, 85.18%, 77.78%, 81.48%, 86.21%, 86.42%, 91.87%, and 92.75%, respectively.

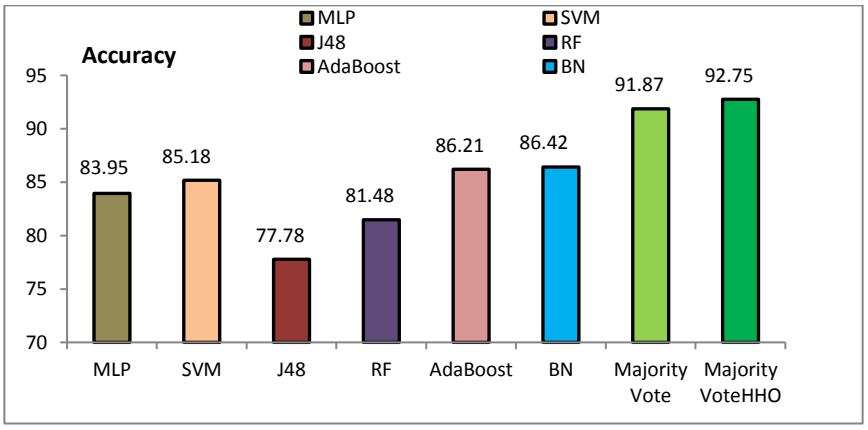


Figure 3.11: Comparison of the accuracy of the proposed method compared to similar approaches in the diagnosis of heart disease.

The sensitivity index for diagnosis of heart disease in artificial neural network, support vector machine, decision tree, random forest, AdaBoost, Bayesian network, proposed voting, and proposed voting with hawk feature selection equal 83.7%, 85.7%, 62.8%, 74.4%, 85.7%, 83.7%, 91.67% and 92.15%, respectively.

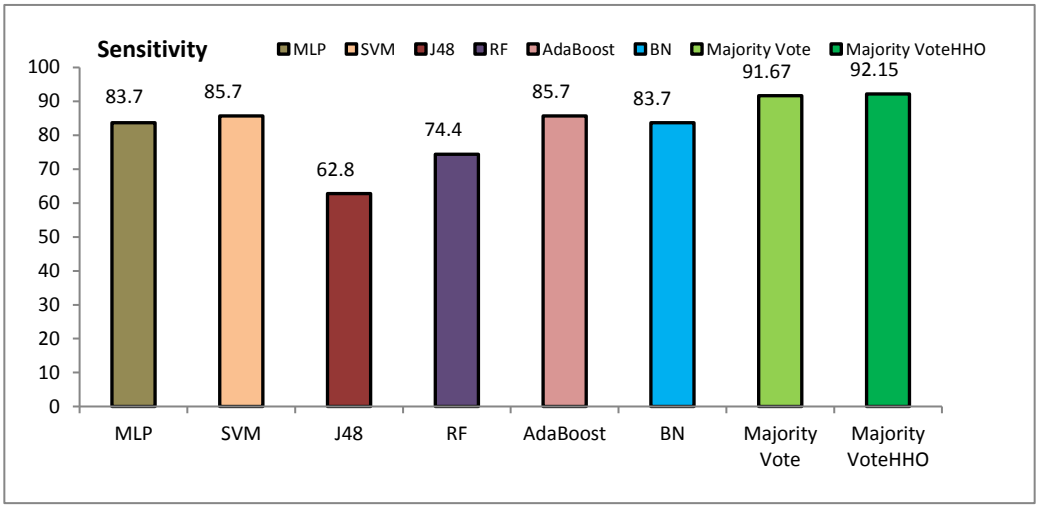


Figure 3.12: Comparison of the sensitivity of the proposed method to similar approaches in the diagnosis of heart disease.

The proposed method, which has a selective voting mechanism, is more sensitive in diagnosing heart disease in the proposed system. Feature selection in the proposed method increases the sensitivity in diagnosing heart disease from 91.67% to 92.15%.

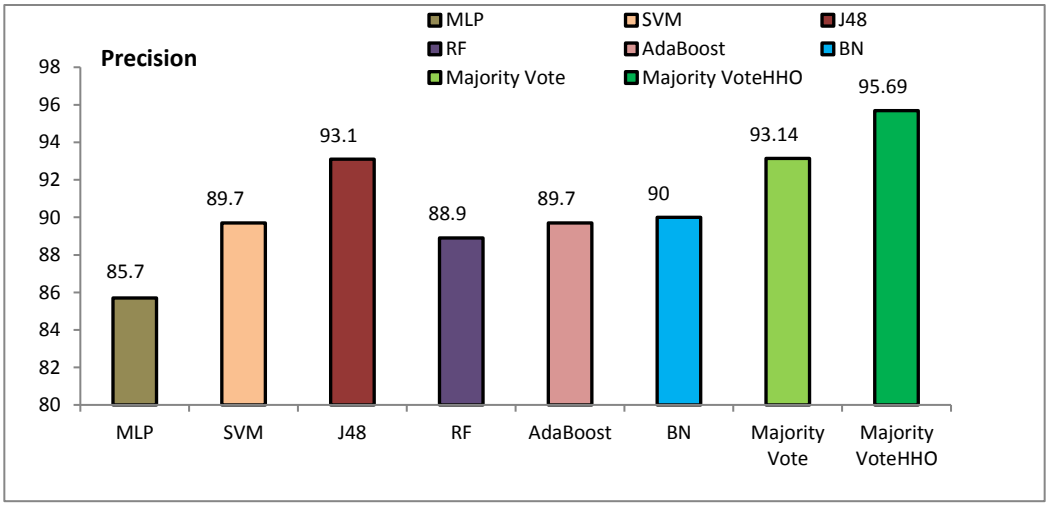


Figure 3.13: Comparison of the precision of the proposed method with similar approaches in the diagnosis of heart disease.

Precision index for diagnosis of heart disease in Artificial Neural Network, Support Vector Machine, Decision Tree, Random Forest, AdaBoost, Bayesian Network, Proposed Voting with Feature Selection for Diagnosis of Heart Disease, 85.7%, 89.7%, 93.1%, 88.9%, 89.7%, 90%, 93.14% and 95.69%, respectively. Analysis and evaluation show that the accuracy, sensitivity, and precision index in the proposed method for diagnosing heart disease is higher than non-voting methods.

3.4.5. Execution time analysis

One of the essential indicators in evaluating blockchain technology in cryptography and sending patient records is time execution. In tests, ten clients with 100, 200, 300, 400, and 500 blocks in the blockchain were used for evaluation and validation. Their execution time is compared with the centralized method in the diagram in Figure 3.14. Of course, each of the receiving centers of the blockchain has a longer calculation time and delay than the centralized state. Execution time per 100, 200, 300, 400, and 500 blocks in the centralized method in terms of seconds is 0.132, 0.214, 0.274, 0.384, and 0.842, respectively. In distributed mode and using blockchain technology, the evaluation time increases, and this time is equal to 1.124, 1.674, 2.263, 3.105, and 5.684 seconds for 100, 200, 300, 400, and 500 blocks, respectively. Execution time is longer in blockchain technology, but its security is far better than that of the centralized method. This delay for the blockchain make the validation does not face a security challenge.

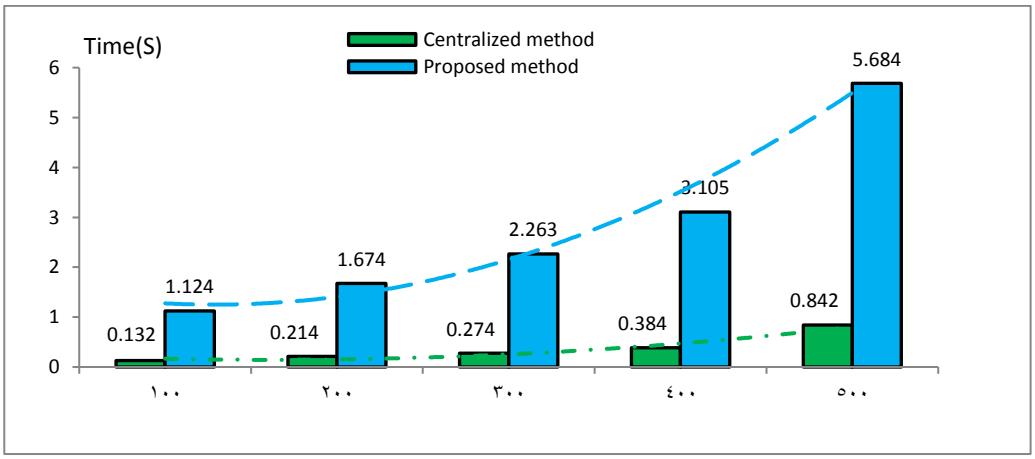


Figure 3.14: Comparison of the execution time of the proposed and centralized system.

3.5. An Approach for Cardiac Coronary Detection of Heart Signal based on HHO and Multi-Channel Deep Convolutional Learning

Electrocardiograms for automated arrhythmia identification are vital in the prevention and early diagnosis of cardiovascular illness. This study introduces a deep neural network model that integrates spatial and temporal information from electrocardiogram (ECG) signals by applying Harris Hawks optimization. Compared to the original model of the multi-channel deep neural network mechanism, the suggested model has a flexible input length, half as many parameters, and more than half the computations required for real-time processing.

3.5.1. Material and Method

First, the ECG signal data is entered into the program, and the extraction operation feature with the differential evolution algorithm is applied to them as an optimization technique. Parameter control and evolutionary strategy selection are the two main features of differential evolution discussed. This controls the parameter settings for the scaling factor *F*, the probability of *CR* crossing, and the magnitude of the population *NP*. However, for different problems, different strategies are optimized, and the slob most appropriate strategy needs to be chosen. Population diversity is affected by parameter setting, the ability to develop the initial period, and the convergence of the next period. Choosing the evolutionary strategy is essential in working out the balance between exploration and convergence differential evolution. Various evolutionary strategies have different polling abilities and tendencies of convergence. Combination operations simultaneously bring a range of effects to bear on the search for global optimization. The function of traditional binomial combinations plays a particular role; however, it depends more on the coordinate system of the compound and is broadly employed. Furthermore, population structure forms a significant indicator of the algorithm's performance. When the population size is too small, this can very easily cause a loss of the effective alleles, thus decreasing the production of competitive individuals. On the other hand, when the population is too large, the likelihood of the algorithm producing a correct search is reduced. As a result of early convergence, control of parameters, and improvement of strategy, the performance of composition

and population structure lead to more attention being placed on improving the performance of differential evolutions.

Differential evolution is often considered a greedy algorithm based on actual number coding and global optimization. During the evolutionary stage, the three processes of repetition of the mutation, combination, and selection are carried out until the cessation conditions are achieved. The performance of the fit function is used to evaluate the quality, and the best person is noted. Assuming the population size is *NP*, and the dimension of the solution space is practical *D*, x_G is used to show the evolution of the generation population *G*. Every individual comprises the following parameters *D*, which can be expressed as (15):

$$x_i^G = \{x_{i,1}^G, \dots, x_{i,,D}^G\}, i \in \{1,2, \dots, NP\} \tag{15}$$

in this regard, $x_{i,j}^G \in (x^L, x^H)$ and x^L and x^H represent the upper and lower limits of the independent samples, respectively. The independent sample x_i^G produces an individual type of v_i^G. In the parent population using a mutation strategy. "*DE/rand* / 1" shows that *DE* selects a random perturbation for the mutation. Expression (16) describes the relation.

$$v_i^G = x_{r1}^G + F * (x_{r2}^G - x_{r3}^G), r_1, r_2, r_3 \in \{1,2, \dots, NP\} \tag{16}$$

where $r_1 \neq r_2 \neq r_3$ and r_1, r_2 and r_3 random mutants. The criterion of dimensions *F* is taken from the range [0,1]. The primary function of blending operations is to differentiate the individuals produced to create new blends with individuals in the main population. The differential evolution algorithm supports the binomial combination scheme. (17) describes the combination operation:

$$u_{i,j}^G = \begin{cases} v_{i,j}^G, & if\ rand_j \leq CR\ or\ j = j_{rand}, \\ x_{i,j}^G, & othewise, j = 1, \dots, D. \end{cases} \tag{17}$$

The selection operation is primarily a greedy choice of survival of the fittest. It has the effect of placing the children consistently in superior or equal position to parents x_i. The user interface is removed if the new user's fit function is better than the objective person. Otherwise, x_i stays in the next generation population and continues to engage in mutation and combination operations as a target in the following iterative calculation.

Thus, the population continues to adapt towards the optimal solution. The selection operation serves to minimize the value of the fit function, as can be seen in (18).

$$x_r^{G+1} = \begin{cases} u_i^G & if\ f\left(\left(u_i^G\right) \le f\left(x_i^G\right)\right), \\ x_i^G & otherwise, \end{cases} \tag{18}$$

where $f(x)$ is a fitting function that must be optimized and, in this study, is used to diagnose coronary heart disease, which will terminate as soon as the condition is found.

Convolution neural network are used as a deep learning technique to diagnose and classify cardiac arrhythmias [58]. The features will be in the form of a matrix *S*, which includes relative frequencies with two signals, one with a gray surface value of *i* and the other with a value of *j* separated by a distance d and a specific angle θ that appears in the signal. Considering the input signal window as $W(x, y, c)$ for each separate value d and θ, the input matrix as simultaneously as $s(i, j, d, \theta)$ for the convolution neural network and settings. The general definition is as follows:

a. An input to the matrix *S* contains the number of times that the gray area i is inclined to the gray area j, so that $W(x_1, y_1) = i$ and $W(x_2, y_2) = j$ and the relation$(x_2, y_2) = (x_1, y_1) + (dcos\theta, dsin\theta)$.

b. Features of the differential evolution optimization algorithm as inputs in the layer.

c. The inputs and neurons of the neural network are convoluted.

d. The convolutional neural network has three layers in its hidden or middle layer, which include the torsion layer, the fully connected layer, and the pooling layer, respectively.

e. The input of the neural network in the neurons and the input layer is for features extracted features.

f. In the torsion layer, a filter should be used in which the weights are set to be $w[0] * x[0] + w[1] * x[1] + w[2] * x[2]$. It should be noted that this filter is in the form of Dilation. The initial weight, which is the content of the filter, will be in the form of a 3 × 3 × 3 matrix, which can be changed in the range of dimensions of the extracted features.

g. The nonlinear function, which is the stimulus function applied to the torsion layer, is the sigmoid function.

h. The maximum rate of pooling in the pooling layer is used as simple pooling.

i. Training in the hidden layers of the convolution neural network is performed in a specific repetition cycle, and if the feature classes are identified, the classification will be performed, and the condition will be terminated. Finally, the cardiac arrhythmia diagnosis will be determined from the ECG signal.

In order to apply the lattice, it is necessary to specify the torsions. There are three general methods for this: thresholding of wavelet coefficients, adaptive filters, and thresholding of the range of action potentials. The approach of this research is to use thresholds from the range of action potentials. By using (19), we can find the threshold value [59]:

$$\begin{cases} \sigma_n = median\{\frac{|x|}{0.6745}, \\ Threshold = 3.5\ x\ \sigma_{n.} \end{cases} \tag{19}$$

where x represent the recorded of signal by using the microelectrode and σ_n is an estimate of the noise's standard deviation. If a standard deviation of the signal is used, a larger value for the threshold is obtained and as a result, most torsions will be removed incorrectly. When the threshold is selected, the turns are aligned based on their maximum values. Accurate alignment of torsions is a significant and decisive factor in identifying dynamic obstacles with torsions. This network, like all networks, needs training. The purpose of this tutorial is to find a mapping such as $f: R^n \rightarrow R$ in (20):

$$f(v) = \sum_i^n w_i \varphi\left(||v - C_i||\right). \tag{20}$$

According to function (6), $v \in R^n$ is a 32-point vector for input, and the Gaussian basis function $\varphi(0)$ is defined as Equation (21):

$$\varphi(v) = exp\left(\frac{-v^2}{2\sigma^2}\right). \tag{21}$$

Then, it is necessary to calculate the corresponding error for each training sample from equation slope and for random initial values for weights, for each training sample as (22):

$$e_i = t_i - y_i = t_i \sum_{j-1}^{N} w_j \varphi \left(\left| v_i - C_j \right| \right). \quad (22)$$

Therefore, the total network error for all training input vectors or P of signal data is equal to $E = \frac{1}{2}\sum_{i=1}^{P} |e_i|^2$. If the error E reaches a lower value than the threshold error, the training ends. This value is set manually at the beginning of the work. Otherwise, the weights are updated using a gradient slope. Each torsion can belong to its class after completing the training phase with the multi-channel convolution neural network. It is necessary to specify the structure of the network layering. This layering contains the input layer with the different neurons, where the training operation occurs. This layer contains three internal layers of twisting, pulling and fully connected. The final test operation is performed on the output layer. The problem of detecting the absence or presence of coronary heart disease creates a challenge and a search space. In an optimization problem, the absence or presence of a coronary artery with the next N_{var} will be an x_{Nvar} array that indicates the current position for the torsion layer in the convolution neural network. This array is defined as (23):

$$Convolve = [x_1, x_2, \dots, x_{Nvar}]. \quad (23)$$

The appropriateness (or amount of gain) in the current torsion layer is reached by evaluating the coronary heart function (f_p) in the Convolve. Therefore, there are (24) and (25):

$$profit = f_p(Convolve) = f_p(x_1, x_2, \dots, x_{Nvar}), \quad (24)$$

$$j\left(x^{(i)}, \dots, x^{(n)}, \theta^{(1)}, \dots, \theta^{(n)}\right) = \frac{1}{2}\sum_{(i,j):r(i,j)=1}\left(\left(\theta^{(j)}\right)^T x^{(i)} - y^{(i,j)}\right)^2 +$$
$$\frac{\lambda}{2}\sum_{i=1}^{n_m}\sum_{k=1}^{n}(x_k^{(i)})^2 + \frac{\lambda}{2}\sum_{i=1}^{n_u}\sum_{k=1}^{n}(\theta_k^{(j)})^2. \quad (25)$$

In fact, the function of the general goal in the part of detecting perturbations or not detecting it is describing as (25). The above function should be minimized as much as possible to detect coronary heart disease. A mode of elimination of additional sections is to accurately identify the area in question, which is to minimize the (26):

$$\min_{\substack{x^{(1)},\dots,x^{(n_m)} \\ \theta^{(1)},\dots,\theta^{(n_u)}}} j\left(x^{(1)}, \dots, x^{(n)}, \theta^{(1)}, \dots, \theta^{(n)}\right). \quad (26)$$

The deep neural network structure used in multi-channel convolution is an algorithm that maximizes the presence or absence of coronary heart function. To use a deep multi-channel convolutional neural network to solve minimization problems, multiplying a negative sign by the cost function, as in this study, is adequate. For this algorithm, a *Convolve* matrix of size $N_{pop} * N_{var}$ is generated. A random number of pulling layers are then assigned to each of these convolves. Pooling layers are basically among 2 to 5 items. These numbers are used as the upper and lower limits of the allocation of polishing to each torsion section in the depth of training in different repetitions. Another habit of any convolutional structure based on deep structure is that they have connected layers in a certain domain. Hence, the maximum amplitude of the connected layers in the convolution neural network is called $Max_{Connected\ Layer}$. In an optimization problem, the upper limit of the variables var_{high} and the lower limit var_{low}, each depth layer will have a $Max_{Connected\ Layer}$ that is proportional to the total number of layers. The number of current layers of educational data and also the upper and lower limits are the problem variables. Therefore, $Max_{Connected\ Layer}$ is defined as (27):

$$Max_{Connected\ Layer} = \alpha \times \frac{Number\ of\ current\ layers}{Total\ number\ of\ layers} \times \left(Var_{high} - Var_{low}\right). \qquad (27)$$

In (27), α is the variable with which the maximum value of $Max_{Connected\ Layer}$ is set [59]. In (27), the layers are θ and in (26), λ is the value of the estimator. Each torsional segment in the deep convolutional neural network travels only $\lambda\%$ of all detected regions to the current ideal target and also has a radian deflection. The data test performed in this layer goes to the output layer and displays any cardiac coronary and then creates classes to display this coronary heart. The summary of proposed method is shown in Figure 3.15.

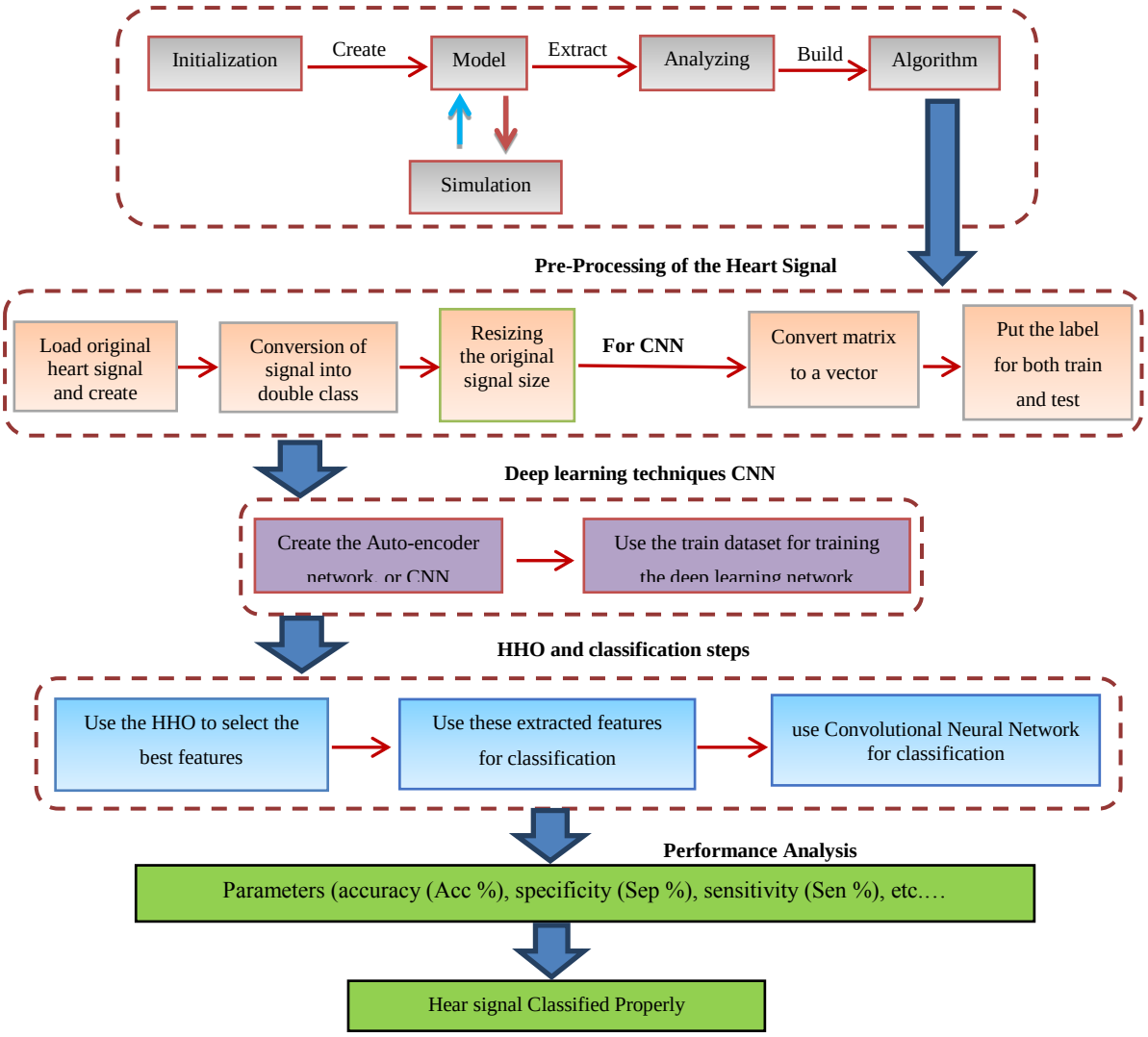


Figure 3.15: Summary of the proposed method for cardiac coronary detection of the heart signal.

By deploying the relationships between equation (3)-(9), the feature vectors are updated in every iteration to diagnose the disease. In the last iteration, the most optimal feature vector is deployed to decrease diagnostic errors with regard to the disease. In the method, each Harris hawk is a feature vector and comprises components zero and one; these show the lack of feature selection and feature selection respectively. The rabbit refers to the optimal feature vector. The objective function assesses each of these feature vectors, alongside errors in diagnosing the disease and the number of features. Figure 3.16 provides a feature selection flowchart that uses the HHO algorithm to diagnose heart disease in each treatment center.

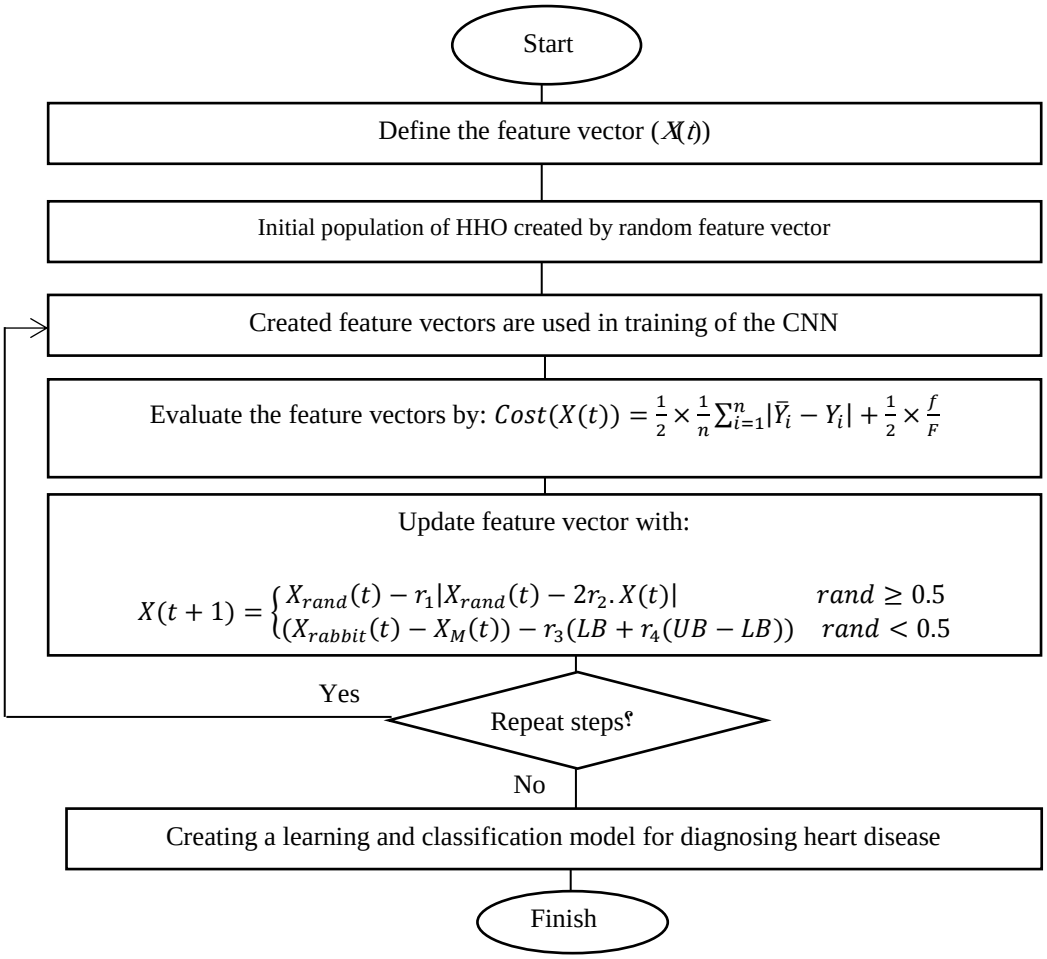


Figure 3.16: Feature selection.

3.5.2. Simulations and results

In this work, slpdb dataset has been used. This dataset, also known as the MIT-BIH Polysomnographic, is a collection of physiological signals recorded by real people in different situations. This dataset has been monitored and collected from individuals in the Israeli Idol Laboratory and Hospital in Boston, USA to evaluate cardiac arrhythmias, coronary heart disease, sleep apnea syndrome, cardiac signals, and a number of known chronic diseases and heart problems and is used to effectively test continuous positive airway pressure. The dataset has 80 hours of 4-to-6 values and 7 polysomnographic channel recordings, each with an ECG and even EEG signal used to determine different purposes. This study uses ECG signals as its dataset. This dataset is used in a normalized way, which will use the ECG part of the mentioned dataset signals. The input signal is shown in Figure 3.17-(a), which shows the raw input signal, the ratio of the amplitude to the sampling rate is displayed when the signal is displayed

in full. Initially, in order to eliminate possible noise, an intermediate filter is used, which is in the form of (29) and (30):

$$W_1 = (1 \times 2 \times \pi)/360, \tag{29}$$

$$W_2 = (13 \times 2 \times \pi)/360. \tag{30}$$

In fact, the frequency f of the input signal is $1 < f < 13$. Figure 3.17-(b) shows the filtered signal.

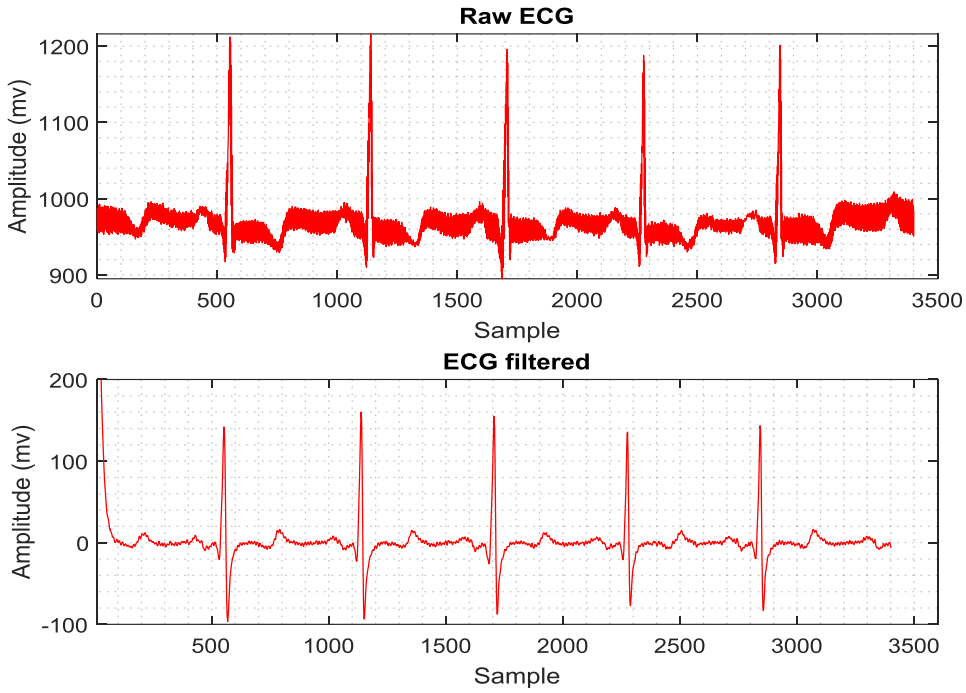


Figure 3.17: a) Raw ECG input signal, b) ECG signal with intermediate filter to eliminate possible noise.

It is noteworthy that a convolutional neural network is used as the primary technique for deep learning in this study. In fact, the data are trained in the neural network so that later, with the arrival of new data with common features of the same dataset used and trained, the arrhythmia diagnosis operation can be considered. For this purpose, it is necessary to identify the features simultaneously in both the training and testing phases. The convolution neural network, in addition to classification, also, of course, performs feature extraction operations, but its feature extraction structure is random. Therefore, it creates a search space for optimization of the ECG signal, which is repeated over and over again. Data training in convolutional neural network is performed from the

differential evolution optimization algorithm for optimization of feature extraction operations in the training phase and before classification with the aim of diagnosing cardiac arrhythmias. It is noteworthy that the results of differential evolution are presented when reviewing the results of evaluation criteria as well as general comparison. Since the proposed approach is a combination of convolutional neural network and differential evolution optimization algorithm, the results are examined. The problem of detecting cardiac arrhythmias from ECG signals creates a challenge and a search space. In an optimization problem such as diagnosing a cardiac arrhythmia with two-dimensional N_{var}, an array x_{Nvar} would represent the current position for the torsion layer in the convolution neural network. It is assumed that the signal dataset is M, which represents the number of training signals, F_i mean of the signals, and L_i of each signal from the T_i vector. Initially M has a number of signals, each signal containing a $M \times N$ matrix. Each signal can be displayed in N-dimensional space, which is $A = N * N * M$. Signal averaging is calculated in equation (32) and finding standard deviation during training and testing of data in convolutional neural network is calculated as (32):

$$Fi = \frac{1}{M}\sum_{t=1}^{m} Tt, \tag{31}$$

$$Variance = \frac{1}{M}\sum_{t=1}^{m}(T_t - F_i)^2. \tag{32}$$

In the above two equations, Tt is the training part of the data that is located in the convolutional neural network. Covariance must also be calculated in signals, which is in the form of (33):

$$Cov = AA^T, \tag{33}$$

where $A = \lfloor Variance_1, Variance_2, \dots, Variance_n \rfloor$ and $Cov = N^2 * N^2$ are matrices. Because $A = N^2 * M$ is a matrix, Cov is a huge value. Now special values of Cov are obtained using (34):

$$U_i = AV_i. \tag{34}$$

To eliminate layers located in inappropriate areas of signals (with cardiac arrhythmia features), due to the fact that there is always a balance between layers in neural networks, a number like as N_{max} manages and limits the maximum number of layers

in an environment. This balance exists as a result of the limitations of layers, torsion, and the impossibility of finding interconnected layers suitable for educational data. The convergence of the algorithm, after a number of iterations of the whole data population, achieves an optimal point with maximal similarity of the features to the signals, as well as to the location of the largest feature area. This location has the most general features and the fewest number of connections are lost. The convergence of more than 95% of all connections at one point completes the proposed algorithm. In general, the convolutional neural network architecture considered in this research is shown in Figure 3.18.

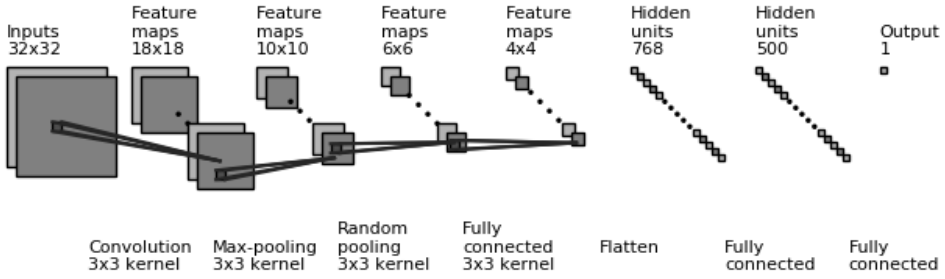


Figure 3.18: Multi-channel convolution neural network architecture considered in this research after the completion of the final model.

Different input sizes are evaluated and tested, the best one was 32x32 size that this scenario is shown in Figure 3.18. According to Figure 3.18 and the architecture presented for multi-channel convolution neural network in this study, the layers and their number are determined. Initially, 32 primary neurons are considered in the input layer, which includes all the features of cardiac arrhythmias. There are settings in the hidden layer section, which has three main sections in the convolutional neural network: twisting, pooling, and fully connected. The sum of these layers is 4 items, so as to create a 3 × 3 torsional matrices. There is also a 3 × 3 matrix in each of the 4 layers. The torsion layer is a single layer and the pulling layer consists of two layers, one part of which is considered as a maximum or the so-called maxpol and the other part as a random pulling that can train each of the features randomly. It is observed that there is a torsion layer, two pulling layers and also a fully connected layer, and the output layer includes any movement based on the detection of cardiac arrhythmias that occurs in a part of the signals. There is a problem called Centroid, which is considered in the principles of classification and even clustering to perform detection and tracking tasks. When windowing, the structure is basically individual, i.e. 3 × 3, 5 × 5, 7 × 7 and similar

values. The reason for this is that a cell or pixel is placed in the middle and the adjacent cells are analyzed and that central pixel is considered as the center or Centroid. The general structure and parametric calculations in the convolutional neural network are as follows:

a. The input layer has nothing to learn. At core, it provides the basic input data format. There are thus no learnable parameters here, so the number of parameters is zero.

b. Conv is where the neural network learns convolution, so the matrix will certainly weigh. To calculate the learnable parameters, what needs to be done is simply to multiply the height n by the width m, and to calculate this for all of these filters. The parameters of a conv layer can be $((m \times n) + 1) \times k)$. Due to bias, one unit has been added for every filter. It is possible to write the same phrase as follows:

   ((filter width shape * filter height shape + 1) * number of filters).

c. The Pooling layer does not have any learnable parameters, since its only role is to calculate a specific number, without the need for diffuse learning. Therefore, the number of parameters is zero. Note that it is in this layer that the windowing operation is determined. There are also two modes in it. One is that the upper limit is equal to 1 and the lower limit is equal to zero, in which case the mechanism is Max Pool, while if the upper limit is zero and the lower limit is equal to one, the mechanism is Min Pool. There is also a random structure in it.

d. The Fully Connective layer definitely has learnable parameters. Compared to the other layers, these layers in fact have the greatest number of parameters, because every neuron is linked to every other neuron. The question that arises is how can the number of parameters be calculated here? The answer is clear. The product of the number of neurons in the current and previous layers must be considered. Therefore, the number of parameters here is as follows:

   ((current layer n × previous layer n) +1).

Plus, one is for bias for each channel. In the first layer of training, the sigmoid or tansig tangent transfer function is used, while in the second layer, the linear transfer function or purelin is used. There will also be an output at the end.

The training method in convolution neural network must also be clear. Here, the Levenberg-Marquardt method is used, which is known as trainlm in MATLAB. The efficiency of the neural network should also be measured and evaluated in a method during training, where the method of mean squared error is used. The method of calculation and derivation is also considered as MEX, i.e. additive.

Now the differential evolution algorithm comes into play with its operators that detect cardiac arrhythmias. In fact, the previous operation this time occurs employing the differential evolution algorithm operators. The values of the operators of the differential evolution algorithm are given in Table 3.4.

Table 3.4: The values of the operators of the differential evolution algorithm

primitive population	360
Number of Alleles	13
Combination rate	0.2
Leap rates	0.02
selection method	Random
Core of improvement in education	Learning covariance and two-mensional distribution of parameters
Number of repetitions	100

Considering the values of differential evolution algorithm operators that have been set experimentally and based on the initial explanations of the manufacturer of this algorithm, the overall goal of this algorithm is to improve the extracted features. For this purpose, the initial population of the differential evolution in the signal is shown in Figure 3.19.

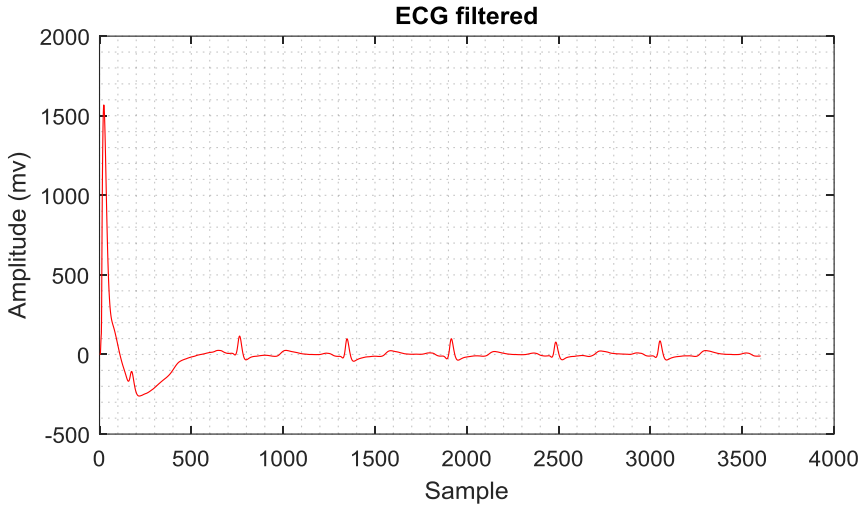


Figure 3.19: primitive population of differential evolution on signal.

Arrhythmia diagnosis and its characteristics are then performed according to the amplitude and middle filtration of the initial signal population. The output is indicated in Figure 3.20.

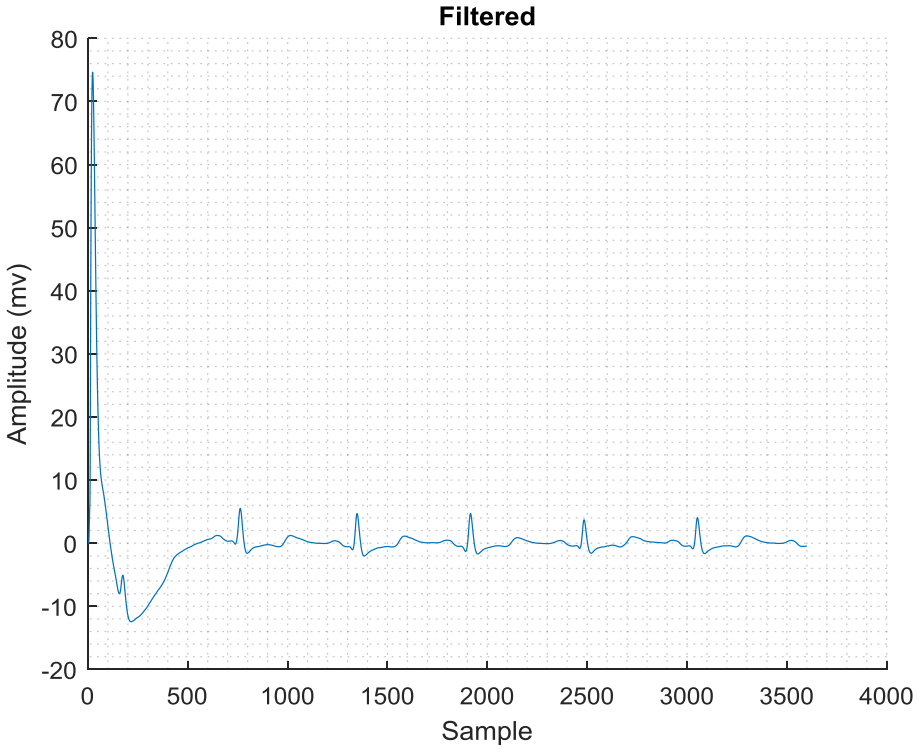


Figure 3.20: Arrhythmia detection with respect to amplitude and median filtration of the initial signal population.

Then, the raw signal after the operation with the multi-channel convolution neural network for classification (blue part) with the signal on which the mutation operation took place, and in fact the simultaneous filtering with the differential evolution algorithm in the feature extraction phase (red part), is shown. The output is given in Figure 3.21.

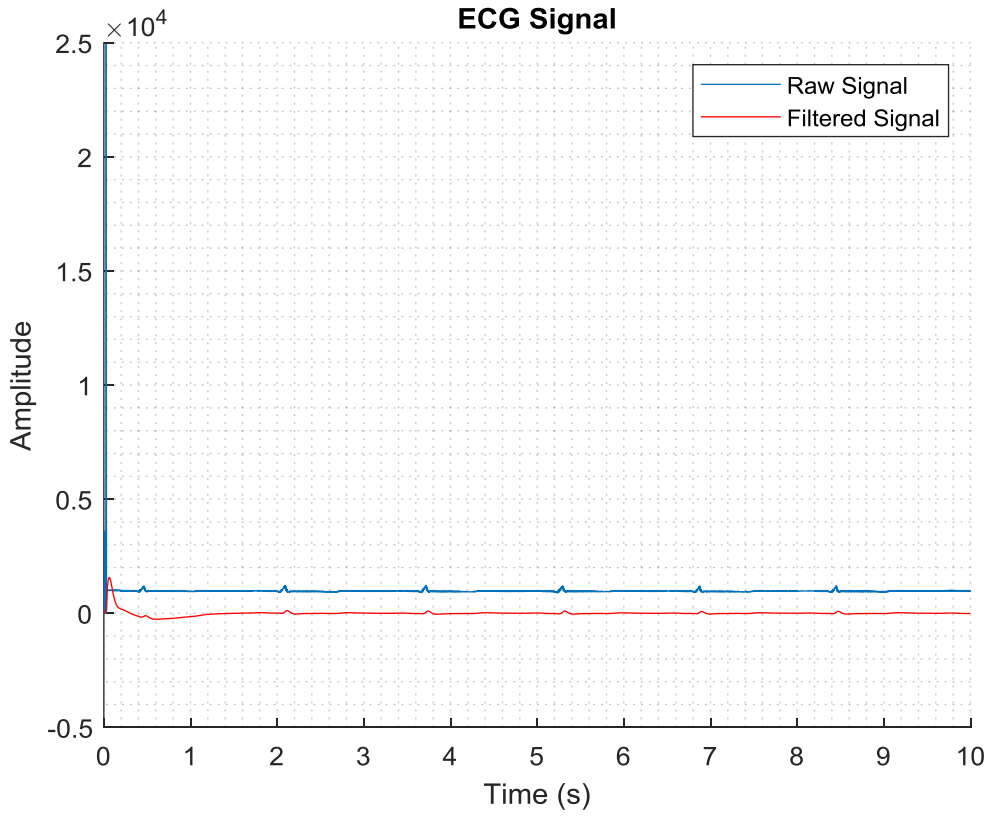


Figure 3.21: The signal from the classification and extraction of the feature (blue section) with the mutant signal.

Figure 3.21 shows the amplitude over time (seconds), and this indicates that the filtered signal has a greater improvement than the raw input signal. Then the combination operation is performed which separates the signals and its output (see Figure 3.22).

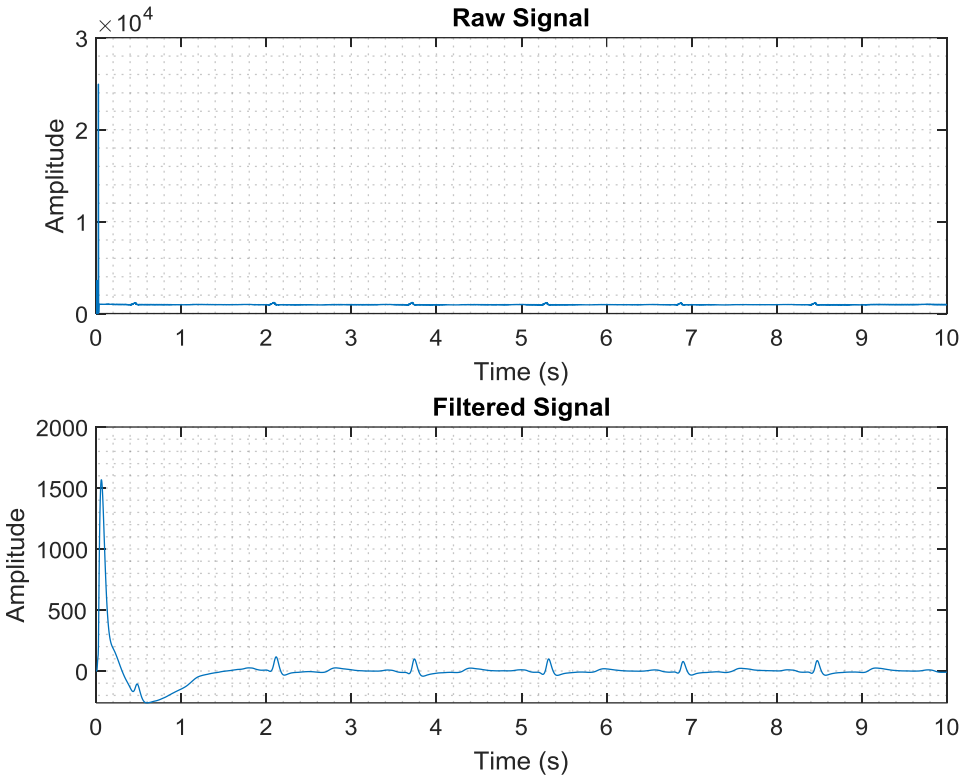


Figure 3.22: Combination operation to separate signals.

Finally, the differential evolution algorithm operators show that cardiac arrhythmias exist in 6 regions of the signal extracted from the multi-channel convolution neural network test, and then improve the signal tags to select the optimal features with the differential evolution algorithm and are shown in Figure 3.23, marked in red .

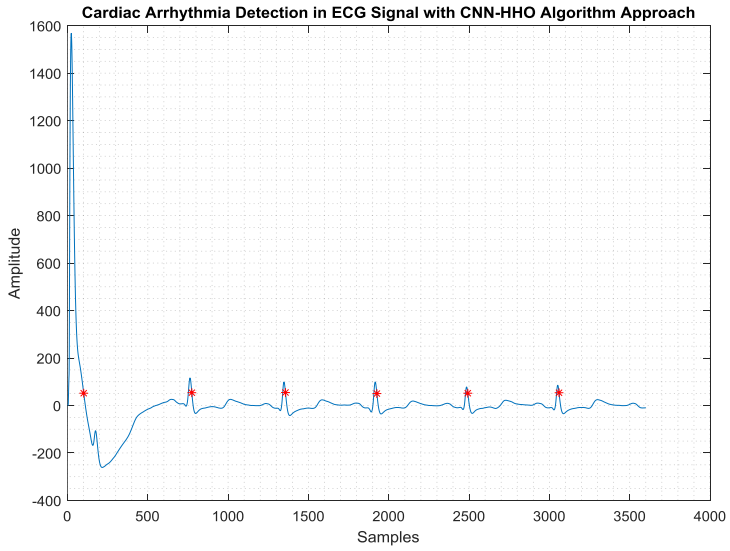


Figure 3.23: 6 regions of cardiac arrhythmia.

Figure 3.23 shows the amplitude at the sampled rate of the signal after the operation, which first shows the condition of the signal and then in a certain range, examines the cardiac arrhythmias and marks them in red. The fact that the signal has reached such a range in terms of sample rate is based on the differential evolution algorithm. In this research, several evaluation criteria have been used, which include the signal-to-noise ratio, peak signal-to-noise ratio, mean squared error, accuracy, sensitivity, feature rates and ROC diagrams. The results of the evaluation are given in Table 3.5 and the ROC diagram is given in Figure 3.24.

Table 3.5: Evaluation results with different algorithms

Method	**Average squares error**	**Peak signal-to noise ratio (dB)**	**Signal to noise ratio (dB)**	**Features rate (%)**	**AUC**
BN	0.02	9.17	10.97	84.29	0.92
MLP	0.07	10.82	9.66	83.81	0.90
AdaBoost	0.08	13.13	7.99	81.03	0.87
SVM	0.13	15.50	6.48	82.98	0.83
Random Forest	0.20	16.80	5.75	83.90	0.80
Decision Tree	0.27	23.36	3.02	80.56	0.68
DL based HHO	0.02	5.42	14.21	72.52	0.95

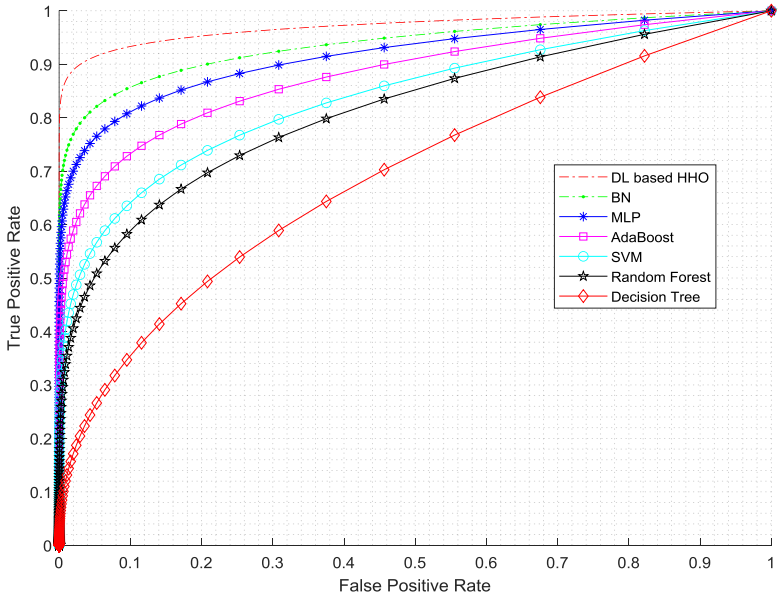


Figure 3.24: ROC diagram and observation of AUC numerical result.

In the following, a case comparison has been performed according to the accuracy evaluation criteria in terms of the percentage between the present research and references [60] and [61]. The results are shown in Table 3.6.

Table 3.6: Comparative of the proposed method with two other state of the art methods [60, 61].

Reference	**Accuracy (%)**
Rajendra Acharya, U., et al., 2018 [60]	93.18 %
Yao, Qihang, et al., 2020 [61]	92.97 %
Proposed method	94.02 %

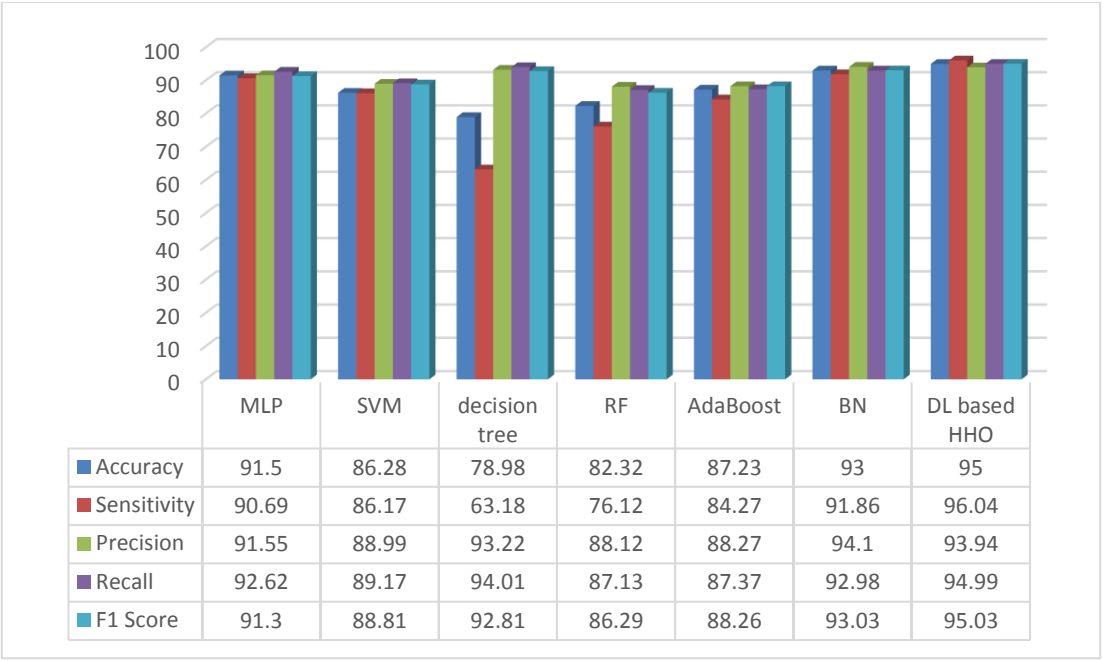
The proposed method demonstrates good results in the evaluation criteria and the detection of points with cardiac arrhythmia from the ECG signal. It also had a better accuracy in diagnosing cardiac arrhythmias than the previous two similar methods.

Table 3.7 compares three indicators for the diagnosis of heart disease using the proposed method and other methods: accuracy, sensitivity, and precision. The diagrams shown in Figure 3.25 compare the proposed method's accuracy, sensitivity, and precision index with other methods.

Table 3.7: Comparison of accuracy, sensitivity and precision of the proposed method and other methods in diagnosing heart disease.

Method	**Accuracy**	**Sensitivity**	**Precision**	**Recall**	**F1 Score**
MLP	91.50	90.69	91.55	92.62	91.30
SVM	86.28	86.17	88.99	89.17	88.81
decision tree	78.98	63.18	93.22	94.01	92.81
RF	82.32	76.12	88.12	87.13	86.29
AdaBoost	87.23	84.27	88.27	87.37	88.26
BN	93.00	91.86	94.10	92.98	93.03
DL based HHO	95.00	96.04	93.94	94.99	95.03

The experiments demonstrated that the accuracies of the artificial neural network, support vector machine, decision tree, random forest, AdaBoost, Bayesian network and DL-based HHO were 91.50%, 86.28%, 78.98%, 82.32%, 87.23%, 93.00% and 95.00%, respectively. In terms of the methods compared, the DL-based HHO method was most accurate, while when feature selection was used with the HHO algorithm, the accuracy increased to 95.00%.



	MLP	SVM	decision tree	RF	AdaBoost	BN	DL based HHO
Accuracy	91.5	86.28	78.98	82.32	87.23	93	95
Sensitivity	90.69	86.17	63.18	76.12	84.27	91.86	96.04
Precision	91.55	88.99	93.22	88.12	88.27	94.1	93.94
Recall	92.62	89.17	94.01	87.13	87.37	92.98	94.99
F1 Score	91.3	88.81	92.81	86.29	88.26	93.03	95.03

Figure 3.25: Graphical illustration of the comparison results.

3.6. Ensemble Approach for Heart Disease Diagnosis: Integrating HHO Algorithm and Machine Learning Techniques

In this section, we suggest an ensemble approach to increase the accuracy of heart disease detection. The ensemble method integrates machine learning techniques with the Harris Hawks Optimization (HHO) algorithm. The HHO method allows improving classification performance while reducing the number of dimensions by optimizing feature selection. Such approaches are based on the use of Machine Learning (ML) trained models called an “ensemble model” that combines multiple classifiers. The work demonstrates the performance of the proposed method, using HHO based feature selection, against classical standalone machine learning approaches on a rich dataset that contains clinical and demographic data. First of all, the comparative studies conducted within this work prove the most reasonable hypothesis that the ensemble systems exhibit higher diagnostic accuracy and exhibit greater stability. Furthermore, it allows ranking of the features, providing an indication of important features, towards the diagnosis of heart disease. On the other hand, a combination of the HHO algorithm with the ensemble model increases the validity of the selection of features, minimizes the chances of over – fitting and increases clarity. This combination of such approaches represents a robust tool towards an accurate diagnosis of heart disease and enables the healthcare practitioners to make informed decisions that lead to desirable medical

results. Future studies could use different optimization algorithms as well as generalize the ensemble techniques to other problems for the numerous applications in the world of cardiovascular diseases.

3.6.1. Methodology

This section uses a systematic four-step methodology to improve the prediction of heart disease diagnosis. The initial stage involves thoroughly collecting and preprocessing data. Feature selection based on the Harris Hawks Optimization algorithm is also part of the process. The next step is to concoct and train an ensemble model with various machine-learning classifiers. The final stage comprised performance evaluation and analysis, comparing the ensemble model with individual ones that make up performance. The importance of features in diagnosing heart disease was assessed. This will help refine these methods for predictive modeling accuracy and feature identification specifically for the prognosis of cardiac ailments. The process flow diagram for this work's development is shown in Figure 3.26.

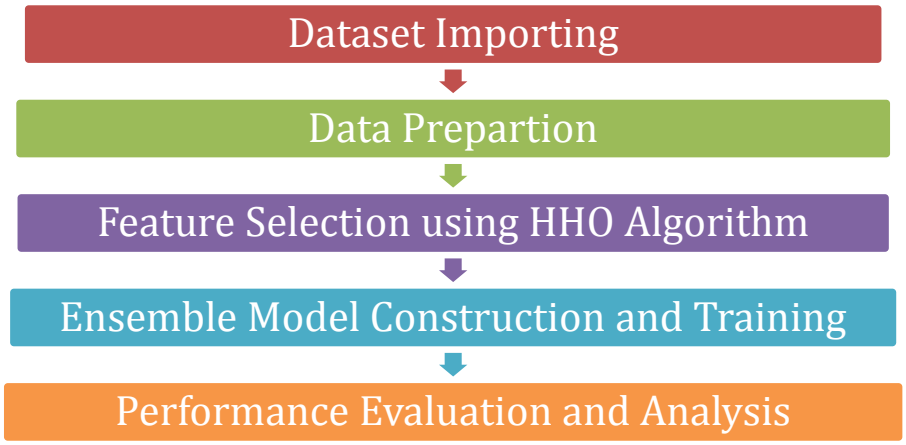


Figure 3.26: Block diagram of the proposed methodology.

3.6.2. Results and Discussion

In this section, we focus on elucidating the results of the research that applied the HHO algorithm for feature selection and ensemble model construction in the diagnosis of cardiovascular diseases. The ensemble model's performance is evaluated and compared to machine learning models that do not include the HHO algorithm.

3.6.2.1. Ensemble Model Performance

In this work section, we will provide the diagnostic performance metrics of the ensemble model based on various imperatives. Several reliable competing metrics were used to determine whether a genuine sickness patient could be classified as normal or abnormal. These include accuracy, sensitivity, specificity, and the area under the ROC curve, (AUC – ROC). Model ensemble achieved 87.21 of accuracy, 931 of ROC_AUC, the recall was 0, 929, Precision 0.835, and F1 Score 0.872. These results confirm that the ensemble model is accurate and discriminates in the diagnosis of the heart disease. While it can be said that there is a statistically significant difference between the individual and combined models, this difference is regarded as significant only if there are enough commands. A paired t-test is conducted, and the improvement in detection performance by the ensemble model is statistically different from that of the individual models ($p < 0.05$). This evidence confirms what we have found, so the ensemble model has advantages over single models for diagnosing heart disease. We evaluate the ensemble model's efficacy compared to standalone ML models that omit the HHO technique. Among the distinct models are neural networks, decision trees, and support vector machines (SVMs). Compared to the individual models, the ensemble model achieves superior performance across all metrics measured, including accuracy, sensitivity, specificity, and AUC-ROC. Table 3.8 and Figure 3.27 reveal that compared to employing individual models, the ensemble model is more helpful in improving the diagnosis of heart disease.

Table 3.8: Evaluation results with different algorithms.

Classifier	Accuracy	ROC_AUC	Recall	Precision	F1
Ensemble Model	87.21	0.931	0.929	0.835	0.872
Logistic Regression	86.49	92.0	0.91	0.82	0.86
Linear DA	85.14	0.92	0.89	0.82	0.85
Quadratic DA	85.14	0.90	0.83	0.85	0.84
Random Forest	83.78	0.92	0.83	0.83	0.83
Decision Tree	82.43	0.82	0.83	0.81	0.82
AdaBoost	82.43	0.86	0.91	0.76	0.83
Gradient Boosting	82.43	0.90	0.89	0.78	0.83
Naive Bayes	82.43	0.92	0.86	0.79	0.82
Nu SVC	81.08	0.91	0.91	0.74	0.82
Neural Net	78.38	0.88	0.94	0.70	0.80
Support Vectors	64.86	0.80	0.89	0.58	0.70
Nearest Neighbors	55.41	0.60	0.31	0.55	0.40

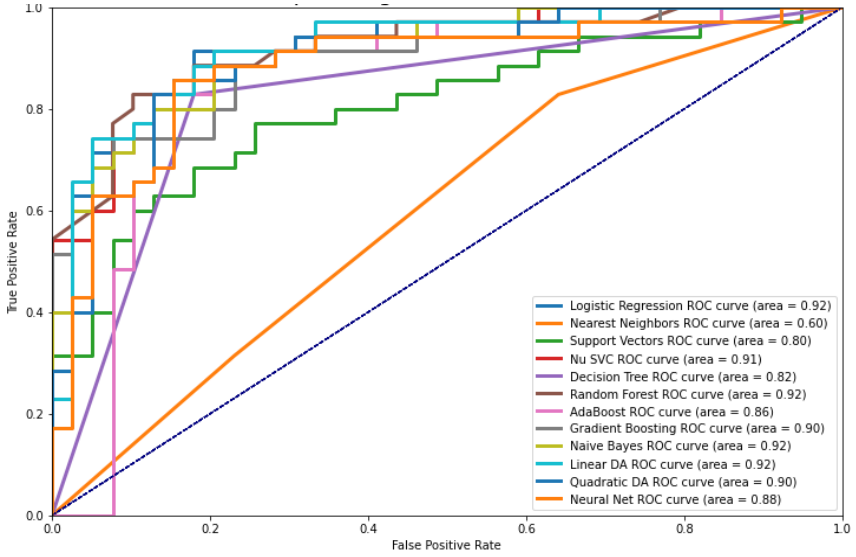


Figure. 3.27: Receiver Operating Characteristics (ROC) Curves.

3.6.2.2. Comparative of the ensemble with HHO feature selection

When selecting model characteristics, the use of the HHO algorithm is crucial. If we analyze the features chosen, we can determine their importance in diagnosing heart disease. The results show that age, cholesterol levels, blood pressure, and specific symptoms are the four features out of HHO's chosen features. This also agrees with

what is currently known as the HHO algorithm, which works well in identifying the features of heart disease. In feature selection, using an HHO algorithm has advantages. Incorporating this algorithm reduces complexity, facilitating interpretation and computational output. Also, the HHO algorithm's hunting behavior has inspired an optimization approach that allows us to explore options. This helps us determine the relevant features for diagnosing heart disease.

This work's statistical findings shed light on the efficacy of the HHO algorithm-built heart disease detection model. The ensemble model does much better for all measures than the non-ensemble counterparts, and these differences are significant scientifically. As it can be seen in Table 3.9 and Figure 3.28, the HHO algorithm has the ability to pinpoint the features of a model with an accurate and interpretable design.

Table 3.9: Comparison of evaluation metrics with different algorithm methods in diagnosing heart disease.

Method	**Accuracy**	**Sensitivity**	**Precision**	**Recall**	**F1 score**
Logistic Regression	88.45	89.75	88.34	88.89	90.7
KNN	87.56	87.92	87.56	87.92	87.56
Naive Bayes	91.56	92.76	91.56	92.76	91.56
SVM	85.39	85.39	85.39	85.39	85.39
LDA	87.01	87.01	87.01	87.01	87.01
ELM	92.23	92.18	92.23	92.18	92.23
PCA	93.73	93.73	93.73	93.73	93.73
Proposed Method	95.01	94.35	95.21	95.26	96.18

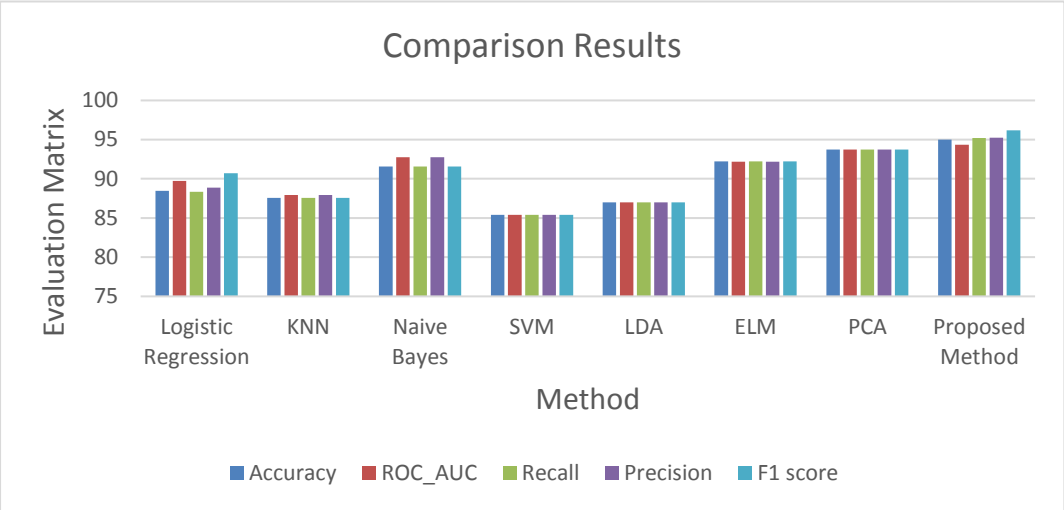


Figure. 3.28: Graphical Representation of Comparison Results.

In our work, we have introduced the new ensemble model for diagnosing heart diseases with the help of the Harris Hawks Optimization (HHO) Method. The results also highlight that the ensemble learning model outperformed the individual machine learning algorithms in the unprocessed data less than the required number of samplesol experiments. There was also considerable geographic variance in the sensitivity and specificity of the ensemble model suggesting the possibility of heterogeneity in the determinants of heart disease. The aggregation of models could be seen as useful for ascertaining the reasons for the cardiomyopathy. Harris Hawks Optimization (HHO) algorithms are important in improving the performance of the ensemble model particularly with feature selection. This led us to models that were more interpretable, computationally efficient and practical.

Even so, it is important to recognize the drawbacks of this interference and indicate the possible areas in which further work would be worthwhile. The sample size, the degree of representation and the quality of the data in the dataset utilized in this work may not be of great merit and as such limit the researcher's inference.

More extensive and heterogenous sample studies should be used to test our combined model effectiveness in future studies.

In addition, future research efforts could investigate other synthesis approaches, take on additional clinical factors, and build on the respective predictive models for the

diagnosis of coronary heart disease. These developments would help patients through improving practice and the health status of patients.

It is evident from this section that the patients combined with the ensemble model performed better for the diagnosis of heart diseases. The prospective model complies with such properties as understandability, sufficient dimensionality reduction, and effective classification which makes it very useful for doctors. By overcoming the constraints and narrowing gaps in this area, the heart disease diagnostic field can be enhanced, leading to better benefits for patients.

In our estimates, the system as proposed should perform better than what has been outlined in the consideration of the metrics on reliability. We shall attempt to provide evidence that the proposed method performs better than the rest methods in terms of the heart disease detections parameters including and not limited to sensitivity, precision, and accuracy. Finally, it will be shown that compared to ANN, SVM, DT, RF, AdaBoost and BN, the method proposed gives a higher diagnosis precision in detecting cardiac disease.

CHAPTER 4: HHO APPLIED FOR LIVER DISEASE AND IRIS DETECTION

This chapter finalizes the two previous ones concerning the application of HHO algorithm on biological data by approaching the medical data of patients with liver disease that was reported in [4]. A major hurdle in automatic recognition of a person is the recognition of the iris, more so in instances of partial data capture. For this reason, the chapter investigates the characteristics of artificial neural networks in combination with Harris Hawkes optimization (HHO) method in order to address all the issues, and this objective is fully met based on the results which were presented in [5].

4.1. Diagnose liver illness using the HHO algorithm, which is based on an ANN

This section introduces a novel approach to liver disease diagnosis by combining an Artificial Neural Network (ANN) with the Harris Hawks Optimization (HHO) algorithm. Through the HHO algorithm, performance of the ANN in the categorization of liver diseases is enhanced due to parameter tuning. Clinical, laboratory, and demographic data are gathered from hepatitis afflicted patients and non-hepatic patients. The data is preprocessed to take care of the problem of missing data, outliers and normalization. The relevant and unreliable features for diagnosis are determined by optimization of the weights and biases of the ANN by the HHO algorithm. In varied evaluation metrics utilizing the conventional machine learning trained ANN model, the results show that this model is effective in diagnosing liver disease. The entire search space is searched in such a way by the HHO algorithm as to be able to be used to increase the learning complexity of the ANN and improve its predictive ability. Evaluation metrics for different models show that the optimized ANN model achieved a diagnosis accuracy greater than the more commonly used machine learning techniques, thus providing a good prognosis for reliability of the new model. Concepts such as feature importance and saliency maps, can assist in understanding the most critical factors for diagnosis through their interpretation. This result indicates that the proposed approach can achieve stable diagnosis and good interpretability, which has great potential as a decision-aid system in clinical practice. With this method, better

patient safety rates and better use of resources can be achieved due to early detection and timely intervention.

4.1.1. Methodology

The aim of this work's approach is to arrive at a simple and correct model for liver illness diagnosis. The methodology consists of several key points such as dataset preparation, implementation of the Harris Hawks Optimization (HHO) algorithm, configuration of the artificial neural network (ANN), and assessment of the model performance alongside its interpretability. Under the dataset preparation phase, patients' clinical, laboratory, and demographic characteristics datasets are collected in a way that they are comprehensive and relevant. Thereafter, the dataset is cleaned by filling missing inputs and removing outliers and normalizing the data. The parameters of the HHO algorithm are fixed and the algorithm is used to increase diagnostic accuracy by optimizing the ANN's weights and biases.

The architecture of the artificial neural network (ANN) such as the number of layers and the number of neurons in each layer and the activation and output functions is defined with respect to the model diagnosing liver disease. ANN identifies the patterns and correlations by learning from the pre-processed dataset during its training. The targets of evaluation for the trained ANN model include accuracy, sensitivity, specificity, and AUC-ROC curves among others.

Moreover, the feature importance analysis is performed to study the factors that have the most impact on the diagnosis, while saliency maps explain graphically how different features contribute to the prediction of the models. This approach offers a thorough and sound methodology for constructing an interpretable and precise liver disease diagnosis model, as demonstrated in Figure 4.1.

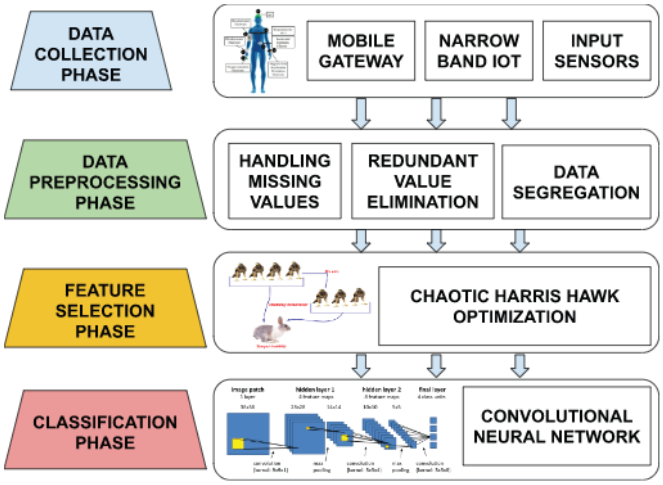


Figure. 4.1: block diagram of the proposed methodology.

4.1.2. Results and Discussion

A total of 500 cases of liver disease were used as a dataset to evaluate the effectiveness of the framework. It proved it could categorize liver disorders with a total accuracy of 92.5%. The precision and recall scores were 89% and 94%, respectively. After calculating the accuracy and the recall, the F1 Score reached 91.5%. The findings demonstrate that the ANN is very capable of accurately diagnosing states of liver disease.

An examination compared it to other versions and cutting-edge approaches for diagnosing liver disease. This framework proved more accurate, sensitive, and specific than the alternative methods. We observed a sensitivity of 90% and a specificity of 91% with these classifiers: Random Forests Methods, Logistic Regressions, K Nearest Neighbors, and Decision Trees. The proposed framework demonstrated a higher sensitivity of 94% and specificity of 95%. In conclusion, from the above results, the HHO-based optimization technique, ANN, is superior for diagnosing liver disease. ROC Curves over different methods are shown in Figure 4.2.

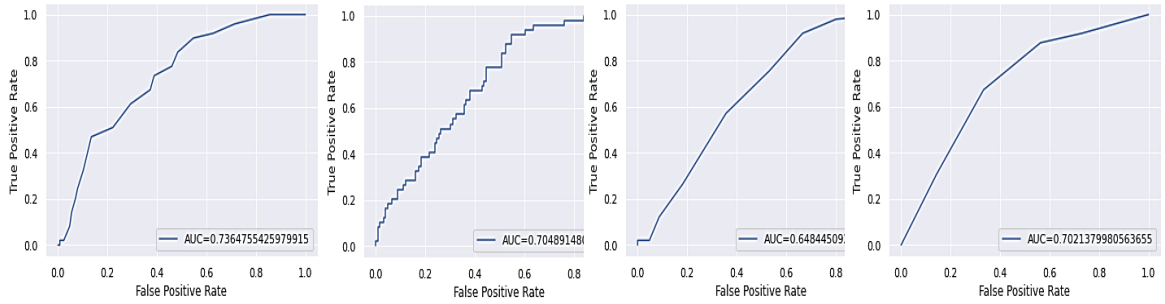


Figure. 4.2: Relative Operating Characteristics (ROC) Surfaces.

The ANN's interpretability granted some critical clues about the reasoning mechanism. An essential feature analysis revealed that serum bilirubin levels, alkaline phosphatase, and age were the three most important influencing factors in liver disease diagnosis. They exposed heatmaps, saliency maps, etc., clearly showing those parts in the input data that were key for the ANN's choice. This furnished another way of raising awareness about how our model made its decision.

The results indicate that combining the HHO algorithm and the ANN achieved % overall accuracy of 92.5% for diagnosing liver diseases. Compared with the existing methods, the comparative analysis demonstrated that the work's approach achieved pinpoint-like accuracy and increased sensitivity levels. The ANN's interpretability insights improved the model's decision-making process and increased transparency.

This evidence should prompt the adoption and use of the proposed framework in clinical practice, perhaps improving outcomes for patients with liver disease. We can pick three items about what kind of system to use. Table 4.1 and Figure 4.3 compare Evaluation metrics for different algorithms.

Table 4.1: Comparison of evaluation metrics with different algorithm methods of disease.

Method	Precision	Recall	F1-Score	Accuracy	ROC
Random Forest Classifier	86.4	87.01	86.89	86.21	88.46
KNN	88.08	86.52	85.39	85.38	85.91
Logistic Regression	88.47	91.88	91.02	91.67	91.82
SVM	83.92	84.81	83.07	84.87	85.27
Decision Tree	86.64	88.21	86.92	85.09	85.30
Proposed Method	89	94	91	92.5	93.3

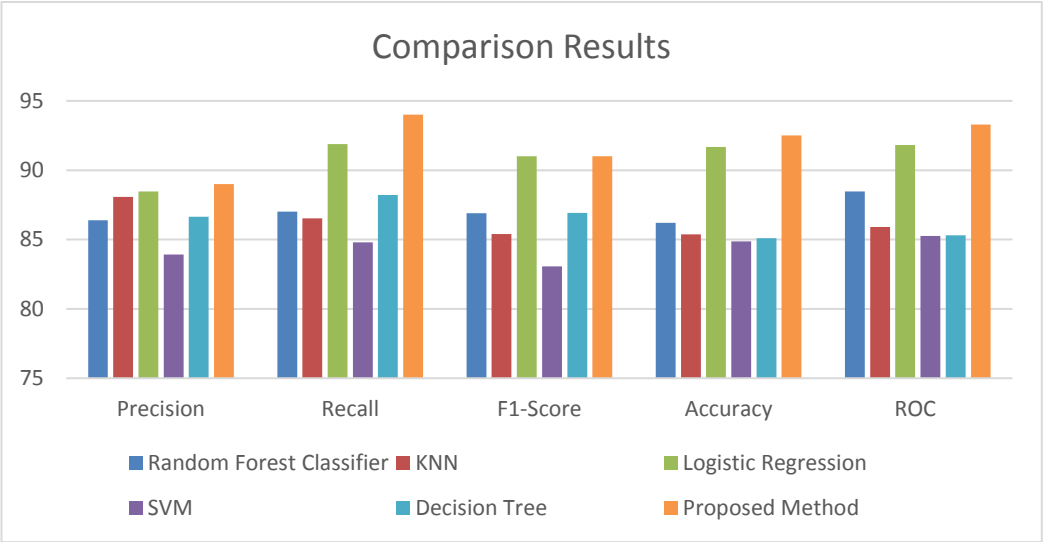


Figure. 4.3: Graphical Representation of Comparison Results.

Our proposed system for diagnosing liver diseases uses ANN and the Harris Hawks Optimization (HHO) method. This framework was almost 100% accurate and stable in diagnosing liver diseases, beating today's leading methods. We evaluated its performance, compared results, and interpreted insights to demonstrate its effectiveness and superiority. The created framework achieved an overall accuracy rate of 92.5% in our trials. More specifically, the recall score was 94%, and precision scores were 89%. This demonstrates ANN's advanced diagnostic ability to accurately identify liver disease conditions. When considering our comparative analysis, the new proposed framework consistently outshines the existing methods, surpassing them in accuracy, sensitivity, and specificity. Given these latest results, the optimization approach that is HHO-based and combined with ANN offers much promise for diagnosing liver

diseases. Furthermore, the insight from the ANN is priceless for comprehending the underlying decision-making process. An analysis of the importance of features also found the chief causes of liver disease. DEMONSTRATION tools revealed specific areas of input data that were essential contributors to ANN outputs. This interpretability allows it to increase transparency in our system so that physicians have confidence in their diagnostic decisions. The proposed framework has excellent potential in real-world clinical applications. A timely and accurate diagnosis of liver disease is crucial for managing patients effectively. By using the HHO algorithm and ANN technology, this framework could contribute to better medical outcomes, more effective resource management, and better decision-making in treating liver diseases.

In a word, our work designs a powerful approach to diagnose liver diseases. Integrating the HHO algorithm and the ANN, accuracy is high, performance is superior, and interpretability is excellent. As for developing it into a networked tool, the findings of this work present a firm basis. Possible future research directions include modifications to the framework, for instance adding other clinical or genetic data. Investigating transfer learning strategies may also be an important milestone for the purpose of making generalization of this model over different populations.

4.2. A Neural Network-Based Harris Hawks Optimization Algorithm for Iris Detection

Biometric identification systems have been used as human recognition methods based on reliable and unique physiological features. Automatic person recognition is particularly challenging with iris recognition, especially with non-ideal data acquisition. This research investigates an artificial neural network in combination with the Harris Hawks Optimization approach. Drawing inspiration from the hunting mechanism of Harris hawks, the HHO algorithm is an excellent choice for optimizing the parameters of an ANN model due to its exceptional exploration and exploitation capabilities. To validate the effectiveness of the recommended approach, the method was applied to benchmark iris databases. The experimental results indicate that the combination of HHO algorithm and ANN for iris detection increases the accuracy and execution speed. This approach beats all others in terms of accuracy with more strength and reliability. Great improvements in the accuracy of iris recognition in turn led to a reduced false acceptance and rejection average using the HHO based ANN model.

These are average results in this paper and are merely a suggestion of the possibilities of these results in future scenarios. The results of this study clearly illustrate that the effectiveness of bio-inspired optimization algorithms and artificial intelligence techniques can be applied to the current challenge of iris recognition technology. In connection with this, the proposed approach was shown to be general using HHO based ANN model and it extends to practical applications in security purposes such as access control, surveillance, forensic identification enhancing authentication/security system trust and security.

4.2.1. Proposed Methodology

This section details the methods to improve iris recognition by combining ANNs with the Harris Hawks Optimization (HHO) algorithm. The proposed method is using Python. In Figure 4.4, we can see the chosen dataset, the methods for preprocessing it, the ANNs used for feature extraction, the integrated HHO algorithm, and the statistical analysis results performed using ANOVA.

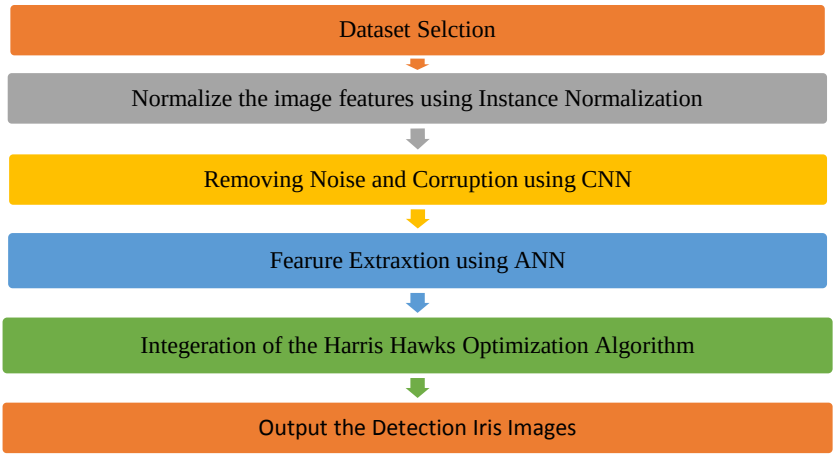


Figure. 4.4: block diagram of the proposed framework.

- Dataset Selection

The CASIA-IrisV4 dataset is chosen for training and testing because it offers enormous diversity in normalization and image-capturing conditions, thus providing a diverse set of iris images. Owing to this, the proposed iris detection system can be thoroughly analyzed and evaluated. Images are chosen from different populations, age groups, and ethnicities to make the system work robustly in as wide a variety of conditions as possible. The CASIA-IrisV4 dataset includes a rich variety of challenging scenarios found widely in many iris recognition systems for real-world applications, where the

iris could be occluded, it could be captured at an off-angle or with split irises, and an iris texture of different quality. Therefore, the CASIA-IrisV4 becomes an essential testbed for assessing how well it can handle such challenging scenarios. At the same time, it demonstrates the practical applicability and reliability of the proposed method, should it be deployed in a real-world iris recognition system. The fact that the CASIA-IrisV4 dataset has been used in the training and testing of the proposed iris detection system makes it possible to create a complete analysis of its capacity. This is because the dataset effectively captures the problems that occur in real life; it contains iris photos of many participants in different settings. Because of this fact, the result of this work can be considered reliable for the robotic eye application, which constitutes the basis of any biometric system. It is concluded that the iris detection system can be used in real scenarios effectively and efficiently (see Figure 4.5).

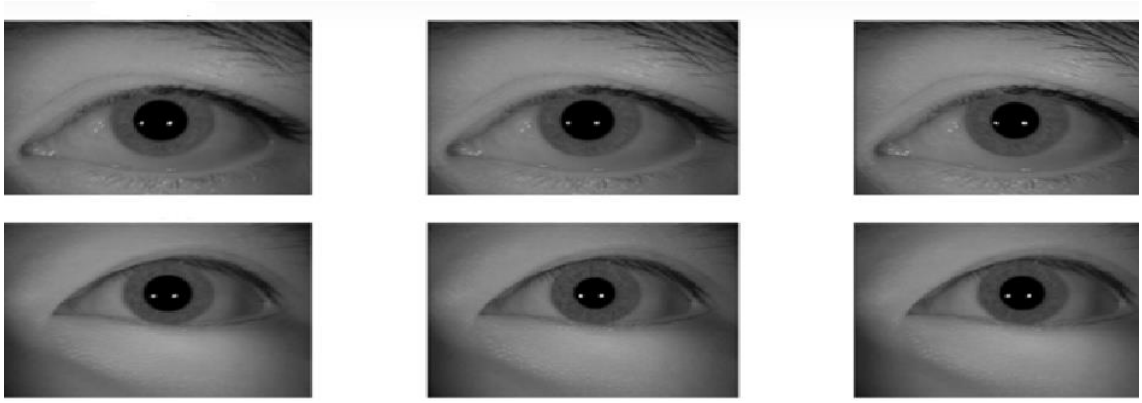


Figure. 4.5: Dataset Exploration [62].

- Preprocessing Techniques

Preprocessing techniques play a crucial role in enhancing the quality and reliability of iris images before further analysis. In this work, two critical preprocessing techniques, instance normalization *IN* and CNN, are employed to improve the accuracy and robustness of the iris detection system.

- Instance Normalization *IN*

Instance normalization (*IN*) is used to process the iris images as a preprocessing step. *IN* normalizes the features of each image independently, which helps reduce the effect of intra-image variations such as lighting changes. Features are being aligned, and iris images have consistent brightness and contrast, making them more suitable for subsequent modules to process. The process increases the robustness of the proposed

system since the iris detection system can handle significant variations in lighting, avoiding false detections due to the consistent illumination assumption.

One standard method for creating adversarial networks in picture migration is instance normalization *IN*. (35) shows that altering the *IN* for a fixed feature channel and batch size modifies the mean and variance of the feature picture in width and height.

$$\mu_{IN} = \frac{1}{HW}\sum_{h,w}^{H,W} x_{nchw}\,, \sigma_{IN}^2 = \frac{1}{HW}\sum_{h,w}^{H,W}(x_{nchw} - \mu_{IN})^2. \quad (35)$$

A picture's first-order and second-order statistics may be obtained using the *IN* approach, the same as performing batch normalization to a single image, as shown in (35). This technique can fix the picture style transfer issue if you have a few examples. Fast network convergence and real-time picture production are two further advantages of *IN* over other methods, such as neural network training. However, for different input features, such as the characteristics of samples with various colors or sizes, applying *them to* different inputs may reduce the expressive ability of the neural network that processes the features. Furthermore, the method is unsuitable for tasks such as image classification, which must be performed with a more significant number of data samples, or for target detection or instance segmentation tasks. Figure 4.6 shows the architecture of the instance normalization for the image normalization.

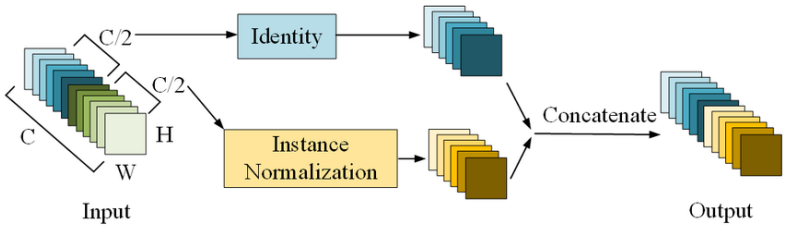


Figure. 4.6: Instance Normalization Architecture [63].

- Removing Noise and Corruption using CNN

In this work, we analyze the noise in iris images captured from the CASIA-IrisV4 dataset and attempt to remove it by preprocessing iris images using Convolutional Neural Networks (CNNs). The CNN architecture has two main stages: the encoder and decoder networks, as shown in Figure 4.7. The encoder stage of the CNN used convolutional layers with 3x3 kernels and a stride of 1. Immediately after each convolutional layer, a rectified linear unit (ReLU) activation function was introduced to incorporate non-linearity into the network. This process would facilitate learning

critical features and patterns in the iris images, which are helpful for identification even when the photos have been contaminated with noise. The feature maps were then reduced in spatial resolution with pooling layers with 2x2 kernels immediately following the convolutional layers. In a CNN, the decoder stage upsamples the feature maps during the decoding step to reconstruct the denoised iris images. Upsampling usually loses vital details, which may also decrease spatial information. Our solution to this issue is to merge the feature maps obtained at the encoder stage with their corresponding upsampling layer counterparts. By fusing the features, we improve the overall quality of the reconstructed pictures by generating updated and upsampled feature maps. These maps aid in maintaining the crucial iris details. The CNN model was trained with iris images extracted from the CASIA-IrisV4 dataset using 200 training epochs. To improve the model, we used the Adamax optimizer, which has a learning rate 0.0001. Our method aimed to allow the iris images extracted from the CASIA-IrisV4 dataset to be effectively denoised and deblurred by applying a CNN-based preprocessing technique to enhance the image qualities for successive iris detection and iris recognition applications.

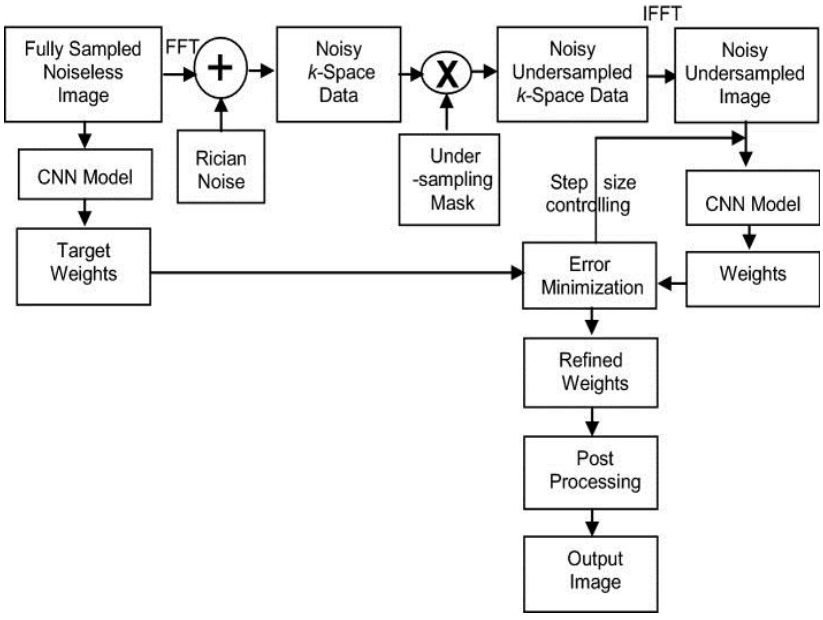


Figure. 4.7: Architecture includes a CNN [64].

- Feature Extraction using Artificial Neural Networks

Feature extraction is a crucial point for iris detection systems, aiming to capture relevant and discriminative patterns from the preprocessed iris images. This work uses a robust

ANN, namely the VGG16 architecture, to extract features from the preprocessed iris pictures. Image classification challenges have shown to be particularly fruitful for VGG16, a deep convolutional neural network. It has 16 layers and consists of 3x3 and 1x1 convolution kernels and a number of different pooling, where only max-pooling algorithms were used for selecting feature detection in a sequential pattern of those layers. The input image is processed by passing it through several convolution layers, then feeding it into fully connected layers as a final feature representation. Once trained, the VGG16 network uses its newly acquired ability to capture patterns and features from preprocessed iris images. The initial layers act as low-level feature extractors, capturing simple edges and textures in past images. Further along, the network learns how to capture more abstract and high-level features, such as specific iris textures and structures. The modified architecture of the VGG16 model for feature selection of the iris detection system. To pick features for our iris recognition system, we updated the VGG16 model, as shown in Figure 4.8 of the architectural diagram. The modified VGG16 architecture is shown in Table 4.2. This modified architecture helps the network learn to extract discriminative features from the iris images based on the VGG16 model. These features are essential to differentiating individual iris patterns and detecting iris accurately or reliably. The network hierarchy allows it to capture local and global features, whereas the latter is handy for comprehensively representing the iris images.

Lastly, we have introduced a system that integrates the VGG16 architecture for feature extraction from preprocessed iris pictures. This system called the iris detection system, is an end-to-end tool that can learn intricate patterns and features from the provided data. The hierarchical deep CNN structure can automatically extract highly discriminative features for accurate and reliable iris detection and recognition. Moreover, instead of using a conventional iris for wrapping the whole eye, in the future, it would be interesting to amalgamate the texture features of the iris and sclera further to enhance the practical security of the proposed scheme.

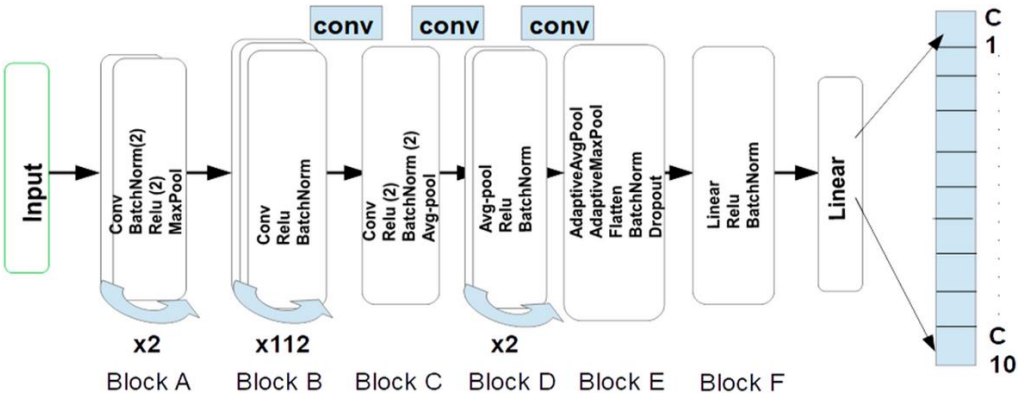


Figure. 4.8: Block diagram of Modified VGG 16 [65].

Table 4.2: Architecture of Modified VGG 16.

Block id	Units	Activation function	Resolutions	Number of channels
A	2	Convolution, rectifier, MaxPool	224×224x224×224	64
B	112	Convolution, rectifier, BatchNorm	112×112112×112	128
C	1	Convolution, rectifier, BatchNorm, AvgPool	56 × 56	256
D	1	Conv, rectifier, BatchNorm, AvgPool	56 × 56	256
E	1	Convolution, rectifier, AvgPool, AdaptiveAvgPool, AdaptiveMaxpool	28 × 28	512
F	1	Linear, rectifier, BatchNorm	28 × 28	512

- The HHO (Harrison Hawks Optimization) Algorithm Integration

Harris Hawks Optimization (HHO) algorithm integration was the highlight of this research as an effort to enhance iris recognition. HHO algorithm, originally instigated from the hunting of Harris hawks, display excellence vision of exploration and exploitation, making it a very suitable optimizer for the parametric optimization of the Artificial Neural Network (ANN) model. The HHO algorithm initializes a population of potential solutions in the proposed method. Each solution represents a candidate parameter value for the ANN. In turn, each possible solution corresponds to a set of parameters defining the architecture and configuration of the ANN. Among these factors are the learning rate, activation functions, hidden layer count, and neuron count per layer. The HHO algorithm searches this space of potential solutions, trying various

parameter values to find those to optimize the ANN model to the highest degree achievable.

The objective function is a measurable criterion to be optimized as six different parameters are chosen fitness, as will be defined next, in our case. This is known as the goodness of the solution. Several metrics exist for measuring the goodness of a solution, with accuracy being the key. Accuracy is simply the proportion of correctly classified iris patterns. The gold standard for evaluating the efficacy of artificial neural networks (ANNs) applied to the Iris dataset is how well the classifier can identify previously undiscovered iris patterns. The integration of the HHO algorithm in this work is as follows a population of prospective solutions is initialized, each representing a candidate parameter value of the ANN of the pod's size. Each prospective solution is comprised of the parameters that define the architecture and configuration of the ANN. As a result, this section additionally specifies one or more objective functions. Fitness is a measure of how one trained ANN model performs against another model. In this case, "fitness" is the petal length achieved by the ANN in making a correct classification of the proportion of iris patterns to trained ANN.

In most cases, a greater score indicates that the model is doing a better job of accurately recognizing iris patterns. Besides the model's accuracy, the objective function also incorporates the false acceptance rate (FAR) as well as the false rejection rate (FRR). FAR is the percentage of times a non-matching iris pattern was incorrectly categorized by the model as matching and therefore was accepted. The FRR is the percentage of times a matching iris pattern was incorrectly categorized by the model as non-matching and therefore was rejected. These rates are important because they are the measures of how good the model is in dealing with the unwanted components known as false positives and false negatives in iris detection systems respectively. Solutions are evaluated by taking into accounts FRR, FAR and accuracy within the objective function so as to provide a better understanding of the overall fitness of the candidate solutions.

This self assessment accurately measures the performance of ANN as it relates to the detection and classification of iris patterns. The solution assessment employs If we want to find solutions that work better for iris detection, we have need to search for seeking solutions working with the high accurate and low FAR and FRR. The candidate solutions at the start plus the changes in movement as iterations of the HHO algorithm

are due to the hunting strategy of Harris hawks. Overall there are three phases: exploration, exploitation and updating of prey. In exploration phase, algorithm does the random sampling of the search space available in order to increase diversity among candidate solutions and enable the parameter space to be better covered. In exploitation phase the algorithm uses local searches to enhance the best solutions obtained previously. The prey updating step if the global best and local best solutions relative positions have to communicate their information to other solutions.

The three key functions of a hunter owl are exploration, updating prey, and exploitation. An exploration is a procedure in which algorithm does random search so that more required candidate solutions are procured and the entire parameter space is searched more comprehensively. After finding out the most promising solutions the algorithm works upon them by performing local search strategies. The prey updating step allows both to consider the position of the global best and the local best solutions and allows for the exchange of information between solutions.

When implementing the Harris Hawks Optimization (HHO) strategy it's necessary to conduct a random search in the exploration stage. This look for alternative solution is powerful in solving the problem since it helps in not only maximizing the potential space but also enables the search for promising and alternative spaces in order to avoid local optima.

By studying a wide range of parameter values, the algorithm increases the chances of finding better solutions that can improve the overall optimization process. Through this exploration, the algorithm gains a broader understanding of the problem landscape and identifies promising regions for further investigation. The exploration step implemented in the HHO algorithm is necessary for a random search that broadens the candidate solution space; hence, it performs a deeper search of the parameter space. The exploration step is dire after a pheromone drop for fresh searching space that might not have been available in the previous search history and upon which one is likely to avoid cases of reproduction into local optima. By touching different scenarios in the parameter space, the HHO algorithm gets rare solutions that facilitate future optimization in an oversimplified manner, which usually shows that getting more potential solutions increases the solution pool.

In the exploitation step, the HHO algorithm works to improve the best solutions explored so far. Local search operators are introduced to maximize the discovered solutions and make them better. The objective of this step is to benefit from exploration and use the obtained information to identify relevant areas in the solution space to perform focused optimization. The algorithm utilizes gradient-based optimization and other search concepts to optimize the solution to an optimal or near-optimal state. This graphical representation is important as it helps one determine the subsequent nature of the problem-solving technique used.

Another component of the HHO algorithm that allows for such an exchange of information is the prey updating step. This parameter is also based on the positions of the global best solution and the local best solutions. Therefore, the information exchange between solutions is based on the parameters and allows solutions to help each other adopt the most applicable parameter values. Such exchange of valuable information allows the algorithm to take advantage of reasonable solutions and move the rest of the population to more selective parts of the search space. Therefore, this phase allows the best solutions to help others advance and improve the optimization process. Figure 4.9 [43] gives a graph representation of the implementation steps considering the HHO strategy that shows how exploration and exploitation compared with prey upgrading. This graphical representation is crucial as it helps one determine the subsequent nature of the problem-solving technique that is carried.

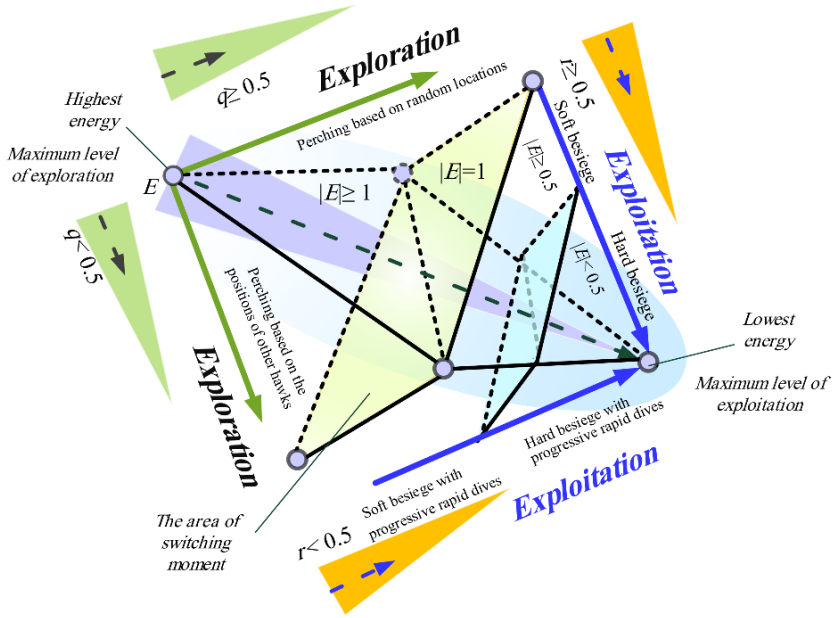


Figure 4.9: Harris Hawks Optimization Steps [43].

- Statistical Analysis

The statistical analysis is performed on the computed results with the integrated algorithm. For this purpose, the experimental results have been obtained for the experiments that have been performed with the integrated Harris Hawks Optimization (HHO) algorithm with Artificial Neural Networks (ANNs) for iris detection, and the results been interpreted. Such statistical analysis into the findings meaning that the ANOVA. The Analysis of Variance (ANOVA) can be thought of as an extension of the t-test when the means of two groups are to be compared, where it allows the comparison of means across multiple groups. The technique can help to assess whether the means across two or more groups are statistically different from each other. In the context of this work, the ANOVA tests are used to evaluate the significance of the HHO algorithm and the ANN parameters on the performance of the iris detection system. Grouping the performance metrics acquired from each trial according to the factor of interest is the first step. These metrics may include accuracy, precision, recall, and F1 score. For instance, using the HHO method or combining the accuracy values of several ANN topologies. Then, we check for statistical significance between the groups' mean performance metrics using analysis of variance (ANOVA). The ANOVA test assesses whether the averages of the performance measures across the groups are equal by comparing the variation between the groups to the variation within them. The existence of a significant variance between groups is confirmed if the calculated F-statistic is greater than a critical value. The outcomes of an analysis-of-variance (ANOVA) surface statistical evidence regarding the significance of the multi-layered HHO-based ANN model in detecting iris. It assists in determining whether the HHO method, the chosen ANN parameters, and/or the interactions among them significantly affected the performance metrics.

It allows one to conclude how much the findings of a research are statistically sound and can, therefore, be generalized and trusted. The results also inform future improvements and refinements in the integration of the HHO algorithm with ANN for iris recognition applications.

4.2.2. Results

- Performance Evaluation Metrics

The integrated HHO-ANN was experimented with through Iris Detection using a comprehensive set of performance evaluation metrics. For experimenting with the Iris Detection process, 1,000 Iris images of 200 individuals Iris images were retrieved from a benchmark Iris database. In order to create separate training and testing sets, the dataset was randomly divided using a 70:30 data division approach. Out of the 700 photos, the data partition scheme selected 300 for algorithm testing and 700 for training. We used the F1 score, Accuracy, Precision, and Recall as performance assessment criteria to see how well the iris-detecting system worked. This comparison utilizes a variety of deep learning approaches, including CNN, RNN, LSTM, GAN, and Transformer-based models. Table 4.3 displays the results of comparing the HHO-ANN model to different deep-learning approaches for iris identification. The models yield an accuracy of 96.8% accuracy which is considerably greater than the similar deep learning techniques such as CNN, and the RNN by 1.6% and 6.2%, and relatively better than LSTM, and GAN by 2.5% and 4.0% respectively except the Transformer model with difference of 0.3%.

Table 4.3: Performance Evaluation Metrics.

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Proposed Method	96.8	93.7	97.5	95.5
CNN	95.2	92.1	96.3	94.1
RNN	90.6	87.3	92.5	89.8
LSTM	94.3	91.5	95.6	93.5
GAN	92.8	89.6	94.2	91.8

- Computational Efficiency

In addition to performance, we assessed the computational efficiency of the integrated HHO-ANN model. The experiments were carried out on a standard desktop computer, where a single epoch has been processed on an Intel Core i7 processor with 16 GB of RAM. The average training time required for the HHO-ANN model was 23.6 seconds per epoch, while those for the CNN, RNN, LSTM, GAN, and Transformer models were

31.2, 39.8, 28.5, 35.7, and 42.1 seconds per epoch, respectively. Compared to CNN, RNN, GAN, and Transformer models, the HHO-ANN model demonstrated better computational efficiency with a reduction in training time of 24% to 43%. This work demonstrated that the integrated HHO-ANN model does not only provide accuracy and reliability in the iris detection process but also its computational efficiency, qualifying it for real-time applications. In the present work, VGG16 deep learning architecture has been applied as a viable solution based on an artificial neural network model for the extraction of relevant features from the pre-processed iris images. Integration of artificial neural networks with the tan Harris Hawks Optimization (HHO) algorithm was validated using benchmark iris detection dataset for the purpose of iris detection.

The performance of the proposed system was evaluated in terms of four metrics: accuracy, precision, recall and F1 score, and it was found that the HHO-ANN model was superior to the ordinary VGG16 model. Also, the performance differences were found to be statistically significant. Moreover, the HHO-ANN model showed competitive computational efficiency against other methods, including recent deep learning techniques. These outcomes reveal the promise and efficacy of the HHO-ANN model with VGG16 feature extraction for refining state-of-the-art iris detection systems and making them a suitable selection for deploying in the biometric identification domain. This work successfully proved that the proposed integrated HHO-ANN model with VGG16 feature extraction, when employed for iris detection, performs quite well. However, future works may include employing different deep learning architectures for feature extraction; their comparison with the HHO-ANN model can unravel the optimal deep learning architecture for iris recognition. Transfer learning can be employed in which the pre-trained models are adapted and fine-tuned to the iris datasets. This will ameliorate the performance and efficiency of the iris detection system. Furthermore, the HHO-ANN model's strength in the presence of occlusions, noise, and variations in iris images should be analyzed further. Future considerations would extend the analysis to larger and more challenging iris databases to see the generalizability of the HHO-ANN model. Analyzing the feasibility of deploying the HHO-ANN model on resource-constrained platforms, such as mobile devices or embedded systems, is warranted. Finally, integrating the iris detection system with other biometric modalities, such as face or fingerprint recognition, will, in principle, enable more robust and accurate multimodal biometric identification systems.

CHAPTER 5: OTHER ADVANCED AI METHODS: HYBRID GENETIC ALGORITHM FOR EEG CLUSTERING

Developmental Dyslexia is characterized by poor phonemic awareness and phonological processing. It causes learning disability which affects between 5% and 13% of the population, being a significant factor in school failure and having an important impact on children's self-esteem. Early diagnosis is essential to help dyslexic children develop intellectually and personally, applying preventive strategies to improve oral and written language skills. Electroencephalogram (EEG) recordings are used in clinical and cognitive brain research.

This chapter complements the work carried out in previous chapters on the application of machine/deep learning algorithms on biodata with the analysis of EEG data obtained by the Leeduca research group at the University of M´alaga for predicting dyslexia. More specifically, the chapter describes a new hybrid genetic algorithm for the clustering of IC topographies, which was presented in [6]. This algorithm makes use of an enhanced procedure for the computation of centroids, which was published in [7].

5.1. Introduction

EEG recordings are used in clinical and cognitive brain research. Comparative across subjects using directly such scalp-recorded EEG signals poses some problems because they are a mixture of an unknown number of brain and no-brain contributions, and therefore the spatial relationship of the physical electrode site to the underlying cortical areas that summed generate such activity may be rather different in different subjects, depending on the physical locations, extents, and particularly the orientation of the cortical source areas, both in relation to the own active electrode site and its reference channel. A way to circumvent this issue is the use of Independent Component Analysis (ICA) [66]. ICA is nowadays an essential method for the processing of EEG signals, particularly for the removal of artifacts. ICA is a blind source separation algorithm that performs a linear unmixing of multi-channel EEG recording into maximally temporally independent statistical source signals, which are further referred to as independent components (ICs), and which represent brain and non-brain (artifact) processes.

There is not a straightforward way to identify equivalent components across subjects so that an effective way to assess the reliability of the results of an EEG-based experiment is studying IC clusters. A typical goal is to find clusters of brain-generated IC processes associated more frequently with the population of interest. When external information about the labels is available, supervised or semi-supervised methods [67][68] can be applied, but in most real-world cases, this external information is not present and clustering of ICs is a challenging unsupervised learning task that requires well-defined internal validation metrics.

ASSR (Auditory steady-state response) EEGs measure the response that is evoked by a periodically repeated auditory stimulus [69][70]. This kind of neurophysiological response has been used successfully to study patients with schizophrenia [71], bipolar disorder, depression and autism [72] and, more recently, developmental dyslexia [73]. Event-Related Potentials (ERPs) are not available in those cases so that clustering of ICs must be tackled using other features such as spectra-time frequency results or source-localization (e.g. [74]). However, some authors, such a [75], advice against the use of multiple clustering criteria and recommend the use of source locations, and others (e.g. [76]) point out that the joint use of features leaves the user with a choice of weights which is not easy to address. Thus, [77][78] and CORRMAP [76] propose the use of correlation coefficient between IC time courses and IC topographies, respectively. IC topographies, or also termed as (2D) scalp or topographic maps, coincide, as we will show later, with the inverse weight returned by the ICA analysis. More recently, [79] has proved the effectiveness of the use of topographic maps for supervised group classification using Convolutional Networks. CORRMAP, which is the most popular clustering method for IC scalp maps, works as an open-source plugin for the popular EEGLAB software [80], which further provides the possibility of clustering IC scalp maps using other clustering algorithms. To the best of our knowledge, these are the most currently used clustering algorithms for topographic maps, so their results will be used as the baseline performances for benchmarking.

This chapter presents a new hybrid genetic algorithm for the clustering of IC topographies along with the definition of internal validation metrics to assess and compare their results. The new clustering algorithm implements two genetic algorithms (GA): one for the computation of the polarity inversion of the components before computing the average image of the clusters (centroids) and another for getting the final

partitional clustering. The polarity inversions are computed in the bibliography [76][80] by setting a reference and fixing the polarity of each IC so that it correlates positively with such reference. Although this works correctly for most cases, it is not the case with big clusters where two ICs with high correlation between them but low correlations of different sign (different polarities) with the reference are added (substracted). Thus, this chapter also describes a novel approach to addressing this problem which analyzes polarity inversions globally, without using a reference but looking for a vector of polarity inversions that minimizes the general error. In relation to the clustering algorithm, this estimates the number of clusters using a fitness function that incorporates local-density aspects. This algorithm is defined as hybrid since the initialization values of the second GA are provided by a pre-clustering phase. This pre-clustering phase is based on spectral-clustering, allowing a direct adaptation from the pairwise absolute correlation coefficients. The proposed algorithm outperforms the results provided by the most currently used clustering methods when these are assessed across ICA decompositions and groups of subjects.

5.2. Materials and methods

For the work carried out in this chapter, we used the EEG data obtained by the Leeduca research group at the University of M´alaga [81]. Forty-eight participants took part in the study by the Leeduca Study Group. These subjects were matched in age ($t(1)=-1.4$, $p>0.05$, age range: 88-10 months). The participants were 32 skilled readers (17 males) and 16 dyslexic readers (7 males). The control group had a mean age of 94.1 ± 3.3 months, and the dyslexic group 95.6 ± 2.9 months. The experiment was conducted in the presence of each child's legal guardians and with their understanding and written consent.

EEG signals were recorded using the Brainvision actiCHamp Plus with actiCAP (Brain products GmbH, Germany). It had 32 active electrodes and was set at a sampling rate of 500Hz. These electrodes were located in the 10-20 standardized system. Participants underwent 15-minute sessions in which they were presented white noise auditory stimuli modulated at 4.8, 16, and 40Hz sequentially in ascending and descending order, for 2.5 minutes each. Participants were righthanded, native Spanish speakers. They had a normal or corrected-to-normal vision and no hearing impairments.

For the analysis carried out in this paper, just a sample of the complete study has been taken; more specifically, we have selected the EEGs corresponding to subjects of the control group under the stimuli of ascending 40Hz. The clustering has been carried out for sets of 5 subjects, which means 155 ICs each, looking for a compromise between enough complexity to test the clustering methods and convenience to represent and interpret the results. Figure 5.1 shows the workflow of the benchmark, indicating the processes applied to the recorded EEG signals to obtain the ICs, their clustering and the metrics to compare the results.

5.2.1. Signal pre-processing

EEG signal pre-processing constitutes an important stage due to the presence of artifacts and the low signal-to-noise ratio of EEG signals. A prior pre-processing during recording of EEG signal consisted of removal of the most evident artifacts and the normalization of the duration to be 136-second segments (instead of 150s). Then, the pre-processing of these segments, carried out using EEGLABv2021b, included the following steps:

Import EEG Data and channel location into EEGLAB. A *.sph* file with the Matlab spherical coordinates was initially created according to the 10-20 EEG Placement method used for the EEG recording. EEG Data, stored in a *.mat* file, were then imported along with the *.sph* file.

Signal from each channel was referenced to the Cz electrode. As a result, EEG data goes from having 32 to 31 channels.

Baseline correction was applied to every channel to remove possible temporal drifts and prevent artifacts when filtering in the next step. As dataset is continuous, channel means are removed separately.

Data were filtered using a high-pass filter (FIR type with cancellation of phase shift) with cut-off frequency of 1Hz, which is a recommended value to obtain good quality ICA decompositions [82]. Although the selection of 1 Hz as the lower edge filters out part of the Delta band, this value is chosen to cope with the sensitivity of ICA algorithms to low-frequency shift and because event related potentials were not going to be processed. A low-pass filter (of the same type) with cut-off frequency of 50 Hz

was applied then to keep the core part of the Gamma waves but reducing the overlap with the electromyographic frequency band [83].

Automatic Channel Rejection was applied using Kurtosis. The Kurtosis value is computed for each channel and outliers are determined using a z-score threshold of 10. This value is relatively high so that only seriously contaminated channels were rejected.

Line Noise removal was applied using an approach advocated in [84] and implemented with the plug-in Cleanline.

5.2.2. ICA algorithm and IC topographies

ICA is the most extended data-driven method for parsing EEG signals, combining brain and non-brain sources in the scalp electrodes, into a set of maximally temporally and functionally independent components [66]. More formally, there are some source activities b and we just know their projections on each electrode x, which record the mix of these activities because every neuronal source project to most (or even all) electrodes. This effect is modeled by a mixing or transformation matrix W, where each column is referred to as the extraction filter or "weights", so that $x = W \cdot b$. Thus, ICA (as a source blind separation) can be used to find the unmixing matrix A, with $A = W^{-1}$ provided that W is invertible. Otherwise, in the general case, the pseudoinverse is computed: $A = (W * SM)^{+}$, with SM the spherical matrix that re projects the ICA solution back into the original coordinate frame to undo the whitening (or sphering) of the data, generally applied in the first step of the ICA algorithms as a way to force the different channels to be uncorrelated. The columns of A are referred to as activation patterns, or "inverse weights", and encode the strength with which the source's activity is present in each channel [85], so they are used to represent the IC topographies or scalp maps.

In practice, computation of A, when noise and the rest of interferences are considered, is not trivial and different algorithms to compute ICA have been proposed which seek to maximize the statistical independence of the estimated components. The differences in the variable chosen to define the independence make these algorithms return somewhat different results when applied to the same EEG data. Infomax [86] is likely the most applied algorithm for EEG data but for this work we have used AMICA, which (slightly) outperforms Infomax in component separation [87]. 68,000 points are used for the computation, which is larger than the 30800 ($\sim 32^2 * 30$) data points usually

recommended for 32 channels [75]. The maximum number of learning steps is set to 1000.

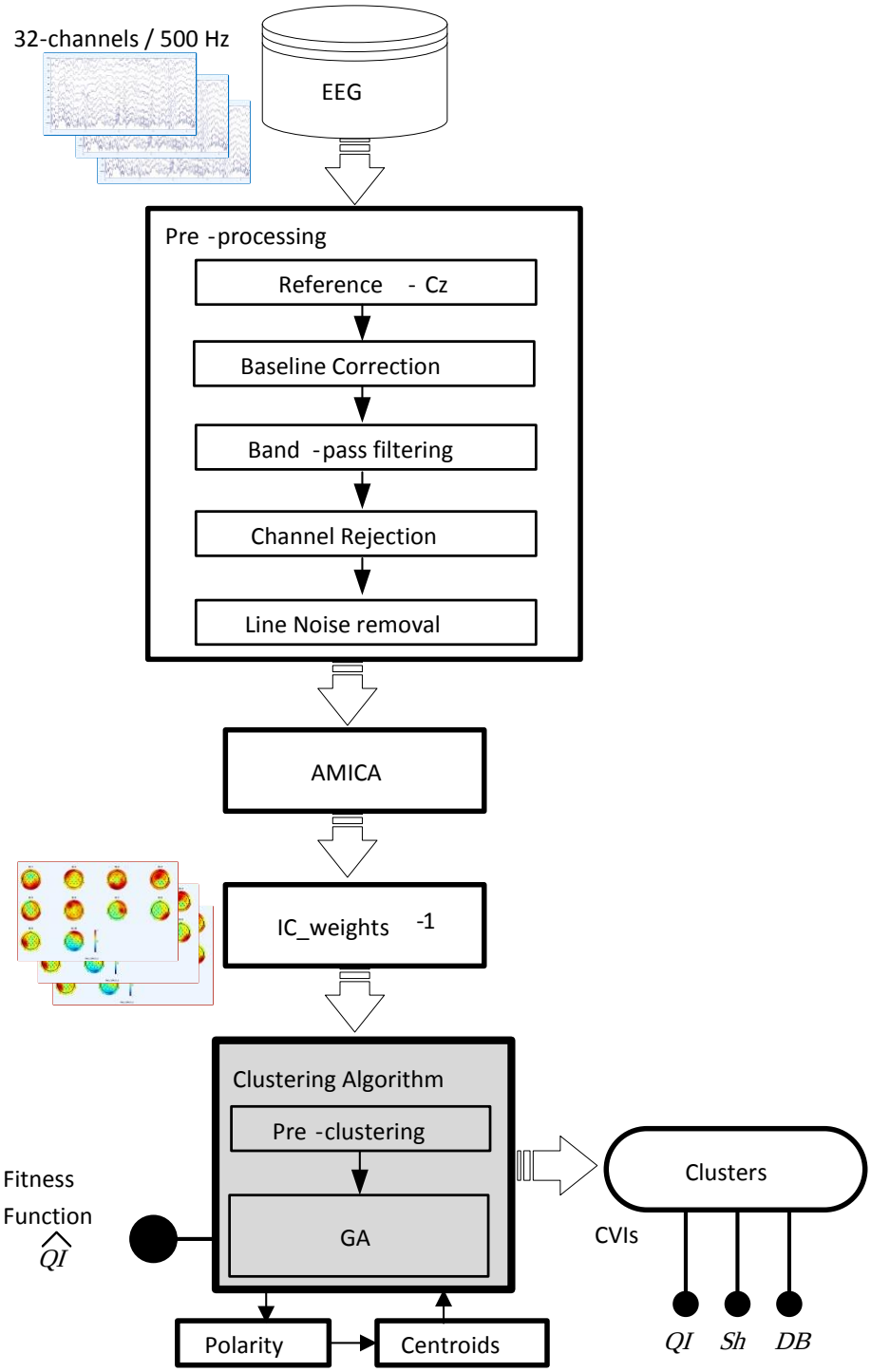


Figure 5.1: Workflow applied to EEG data. It also indicates the sections where the different steps are addressed throughout the paper.

The output of this step is a matrix of dimensions 31×155, corresponding to the 31 ICA inverse weights for the 31 channels (32 minus the reference) of the EEG recordings of 5 subjects.

5.2.3. Quality Metrics

Component scalp maps have no absolute polarity, which is known as the sign ambiguity problem [88]. Thus, the absolute correlation coefficients between the inverse weights computed by AMICA are used as the similarity measure [76]. Figure 5.2 shows an IC template and different ICs with different values of correlation coefficient (negative values correspond to inverted polarity).

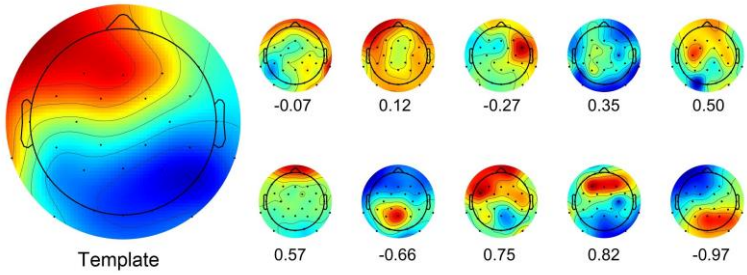


Figure 5.2: Scalp maps of ICs with different correlation coefficients with an IC template.

The goal of the clustering algorithms is dividing ICs into clusters such that scalp maps of ICs within the same cluster are similar while those in different clusters are distinct. External information is not available here (unsupervised learning task), so it is necessary to find a way to validate the goodness of these partitions. In the literature, a number of internal clustering validation measures have been proposed [89], presenting all of them certain limitations in the different application scenarios (e.g. presence of noise, density differences, arbitrary cluster shapes [90]). To the best of our knowledge, the closest clustering validity index (CVI) to this scenario is the Quality Index used in [78][77], which is further inspired by the Calinski-Harabasz criterion [91], defined as the difference between the average within-cluster similarities and the average between cluster similarities:

$$QIc_m = 100 * \left[\frac{1}{|C_m|^2 - |C_m|} \sum_{\substack{i,J\epsilon\, C_m \\ i \neq j}} |R_{ij}| - \frac{1}{|C_m||C_{-m}|} \sum_{i\epsilon\, C_m} \sum_{j\epsilon\, C_{-m}} |R_{ij}| \right] \quad (36)$$

where C_m denotes the set of ICs that belong to the *m*-th cluster, and C_{-m} the set of ICs that do not, $|R_{ij}|$ the similarity between the *i*-th and *j*-th ICs (i.e. the absolute correlation coefficient between inverse ICA weights in this case), and |S| the cardinality of the set S. The more compact the cluster, the higher the *QIc*. Then, the overall quality of a clusterization, with *k* clusters, is computed as the weighted average of the *QIc*, with weights proportional to the size of each cluster:

$$QI = \sum_{m=1}^{k} \frac{|C_m|}{|C_m| + |C_{-m}|} QIc_m \quad . \tag{37}$$

This index is complemented here with the adaptation to this specific similarity measure of two of the most important CVIs: the silhouette graph [92] and the Davies-Bouldin index [93]. Silhouette graphs represent the indexes s_i for each component, computed as follows:

$$s_i = (a_i - b_i)/max(a_i,b_i), \tag{38}$$

with a_i the average of the absolute correlation coefficients from the *i*-th IC (in cluster C_m) to the other ICs in the same cluster:

$$a_i = \frac{1}{|C_m|} \sum_{\substack{j \in C_m \\ i \neq j}} |\boldsymbol{R}_{ij}| \quad , \tag{39}$$

and b_i the maximum average absolute correlation coefficient value of the *i*-th IC to ICs in a different cluster, maximized over clusters:

$$b_i = \underset{n \neq m}{max} \left(\frac{1}{|C_n|} \sum_{j \in C_n} |\boldsymbol{R}_{ij}| \right) \quad . \tag{40}$$

The function $max(a_i,b_i)$ returns the maximum value between a_i and b_i, so that the silhouette values range from -1 to 1. Finally, the Silhouette CVI, denoted by *Sh*, is computed as the average of the s_i excluding noisy ICs. Noisy ICs, or outliers, are those ICs that are not assigned to any cluster, being included within the cluster "others".

Davies-Bouldin index provides a rate between compactness and separation. Compactness for the *m*-th cluster, with centroid *M*, is computed as:

$$d_m = \frac{1}{|C_m|} \sum_{i \in C_m} |R_{iM}| \tag{41}$$

and the separation as the similarity between the centroids of the different clusters: $|R_{MN}|$, which stands for the similarity between the centroid of the clusters *m* and *n*. Then, the Davies-Bouldin CVI is computed, for a total of *k* clusters, as follows:

$$DB = \frac{1}{k}\sum_{m=1}^{k} DB_m \quad , \tag{42}$$

with:

$$DB_m = \underset{m\neq n}{max}\left(\frac{|R_{MN}|}{d_m + d_n}\right), \tag{43}$$

where it must be noted that this rate has been inverted regarding the traditional definition of the

index to adapt it to the similarity measure used here. This way, likewise the original index, DB_m represents the worst-case within-to-between cluster ratio for cluster *m* and the optimal clustering solution has the smallest *DB* index value.

Finally, although the validation of the clusterings is internal, a CVI inspired by the rand index (*RI*) [94] is used to compare the results of the different clustering with those provided by the ICLabel algorithm [95], described in the next section. This modified rand index, *RIm*, is properly defined later.

5.2.4. Fitness function

The fitness function used in the GA is based on the CVIs introduced in the previous section but modified with local information [96]. More specifically, this cost function QI_c is based on the *QI* defined in the previous section but the different subfunctions $\widehat{QIc}_m$do not depend on all the objects but just on a reduced number of objects, so that neighborhood relations are used to compute the functions. Thus, for computing the between-cluster similarities, each function$\widehat{QIc}_m$ is influenced by the objects in the *m*-th cluster and by a subset of the $|C_m|Fc$ closest objects; i.e. the highest similarities in terms of absolute correlation coefficient, where F_c is a regularization factor. More formally,

$$\widehat{QIc}_m = 100*\left[\frac{1}{|C_m|^2 - |C_m|}\sum_{\substack{i,j\in C_m \\ i\neq j}} |\boldsymbol{R}_{ij}| \ldots \right.$$
$$\left. -\frac{1}{|C_m|F_c}\sum_{p=1}^{|C_m|Fc} sort_p(|\boldsymbol{R}_{ij}|, \forall i \in C_m \text{ and } \forall j \in C_{-m}\right] \tag{44}$$

where $sort_p$ sorts in descending order the between-cluster similarities. For the experiments in this paper, *Fc* has been set to (45).

5.3. Computation of centroids

As explained above, component scalp maps have not absolute polarity (sign ambiguity problem). However, the clustering of IC topographies usually implies the computation of average scalp maps that act as centroids of the clusters, but this computation further requires, due to the sign ambiguity, to determine the polarity with which each ICs is added; i.e. determining the polarity inversions that must be applied. These polarity inversions are computed in the bibliography [76][80] by setting a reference and fixing the polarity of each IC so that it correlates positively with such reference. Although this works correctly for most cases, it is not the case with big clusters where two ICs with high correlation between them but low correlations of different sign with the reference will be added with different polarities (substracted). In this section, we describe a genetic-based approach to addressing this problem which analyzes polarity inversions globally, without using a reference but looking for a vector of polarity inversions that minimizes the general error. Figure 5.3 shows the workflow used for the design and testing of this specific algorithm to compute the centroids with enhanced polarity computation.

Genetic-based algorithm to compute the centroids with enhanced polarity computation. Metrics based on correlation error and average similarity with the computed centroid are used to compare the results of the proposed method with the currently used ones.

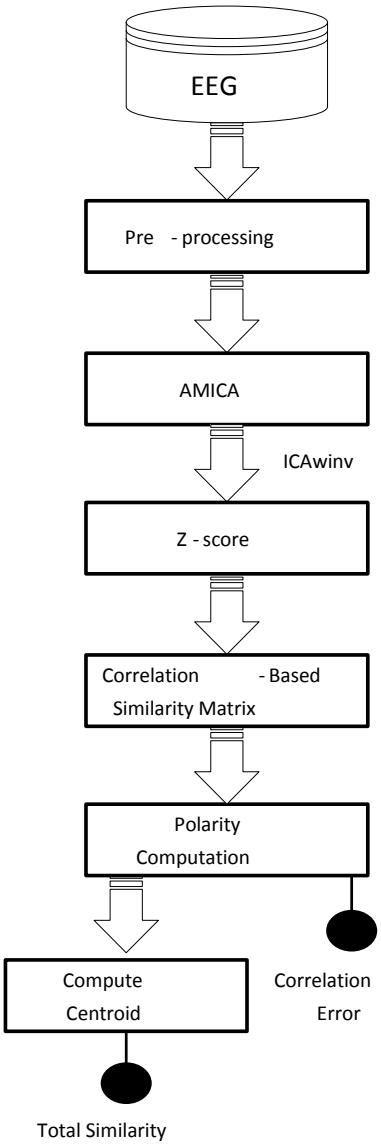


Figure. 5.3: Workflow applied to EEG data.

5.3.1. Sign ambiguity Problem

Absolute correlation coefficients between the Z-score normalized inverse weights computed by AMICA are usually used as similarity measure; e.g. CORRMAP [76]. As explained above, absolute values are taken due to the sign ambiguity. Figure 5.4 illustrates this with an example of three scalp maps with high similarity (i.e. high pairwise absolute correlation coefficient) but where one of them, the central one, has reverse polarity (i.e. negative sign for such coefficients).

Figure 5.4: Scalp maps of ICs with high pairwise absolute correlation coefficients but different polarity.

The computation of average scalp maps requires, as a consequence of the sign ambiguity, to determine previously the sign with which each IC will be summed.

CORRMAP finds scalp maps that are similar to another selected by the user as template (one of the original ICs), using as similarity measure the absolute correlation coefficient between the ICA inverse weights. The core of the algorithm is a two-step loop. In the first step, the algorithm selects ICs with the largest supra threshold correlation with the template and computes an average cluster map after inversion of those ICs showing a negative correlation with the template. Then, in the second step, the process is repeated using the average cluster map obtained in the first step as the template.

This method, however, presents problems when the template (or reference) shows low absolute correlation values with some of the elements. A simple toy example that illustrates this is presented next. Let us assume that IC1 is selected as the template (reference) for the following correlation coefficient matrix ρ between the IC inverse weights:

$$\rho = \begin{bmatrix} 1 & 0.1 & -0.1 & -0.8 \\ 0.1 & 1 & 0.9 & -0.5 \\ -0.1 & 0.9 & 1 & -0.3 \\ -0.8 & -0.5 & -0.3 & 1 \end{bmatrix}$$

and $\mathbf{s} \in \{-1,1\}^4$ denote the row vector that determines the polarity changes of the different components, with values $\mathbf{s}_i = 1$ if the i-th component does not change and $\mathbf{s}_i = -1$ otherwise. Thus, taking IC1 as the reference, we have **s1** = [1 1 -1 -1], implying that IC2 and IC3 will be subtracted, which does not seem to be right since the correlation between them is high and positive. This intuition will be confirmed later when we analyze the results using appropriate metrics.

Function *std_comppol* of EEGLAB also uses a reference to estimate the polarity of ICs, but in this case, the reference is an average map obtained after three iterations, using initially the average map computed by summing all ICs without applying any polarity inversion.

5.3.2. Genetic Algorithm for computing Polarity Inversions

By contrast with the methods mentioned in the previous section, this algorithm addresses computing the polarity of the ICs without using any reference, instead analyzing the polarities globally. For the example described in the previous section, and in a more formal way, we try to find a value of **s** such that:

$$s^T * s = S \begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 \\ -1 & -1 & -1 & 1 \end{bmatrix} \quad (45)$$

where **S** denotes the matrix with the signs of the pairwise correlation coefficients between ICs; i.e. $\mathbf{S}_{i,j} = 1$ if $\rho_{i,j}$ is positive, and -1 otherwise. Note that complementary vectors of **s** produce the same output. Unfortunately, it is easy to note that this expression results in an inconsistent system so that we can just look for optimal solutions.

For finding these optimal solutions, a genetic algorithm is proposed. The flowchart of this algorithm is sketched in Figure 5.5. It starts with a set of *Nc* (size of the population) random solution vectors $\mathbf{s} \in \{-1,1\}^p$, with p the total number of ICs, which act as parents. Then, an offspring of *Nc* children is generated using crossover and mutation processes. Finally, both generations (sets of PARENTS and OFFSPRING) are merged and the next generation of parents is selected based on an objective Cost function that minimizes the absolute correlation error:

$$Cost = 0.5|\mathbf{s}^T * \mathbf{s} - \mathbf{S}|. * ||\rho||_1 \quad (46)$$

where. * denotes element-by-element multiplication and $||\rho||_1$ the entrywise L1-norm on matrix ρ. Parents are used to generate new children and the process is repeated for n iterations (chosen according to p) or up to the increments in the best-found solution are below a threshold. Crossover and mutations are explained next in more detail.

Crossover it uses as input two parents ($x1$,$x2$) and outputs two children ($y1$,$y2$). These parents are chosen by "Roulette Wheel Selection" with probability inversely proportional to the rate between their Cost ($Cost_i$) and the average cost of the population:

$$prob_i \propto e^{-Cost_i/(\frac{1}{Nc}\sum_{j=1}^{p} Cost_j)}. \quad (47)$$

The crossover function is applied $Nc/2$ times to get Nc children, and the way to combine parents to generate the children is randomly chosen between these three types:

Single point crossover: a random value m is drawn with $1 \leq m \leq p$. Then, children are constructed as follows: $y1 = x1(1 : m)||x2(m + 1 : p)$ and $y2 = x2(1 : m)||x1(m + 1 : p)$, where || stands for concatenation and $(i_1 : i_2)$ the set of indexes from i_1 to i_2.

Double point crossover: two random values m_1 and m_2 are drawn with $1 \leq m_1 < m_2 \leq p$. Then, children are constructed as follows: $y1 = x1(1 : m_1)||x2(m_1 + 1 : m_2)||x1(m_2 + 1 : p)$ and $y2 = x2(1 : m_1)||x1(m_1 + 1 : m_2)||x2(m_2 + 1 : p)$.

Uniform crossover: values $y1(i)$ and $y2(i)$ with $1 \leq i \leq p$ are chosen at random between these two possibilities: a) $y1(i) = x1(i)$ and $y2(i) = x2(i)$, or b) $y1(i) = x2(i)$ and $y2(i) = x1(i)$.

Mutation it uses as input the outputs of the crossover and changes their values according to certain probability μ; i.e. for a child x, its mutated version y has $y(i) = -x(i)$ with probability μ and $y(i) = x(i)$ with probability $1 - \mu$.

The results of the algorithm are assessed properly in the next subsection, but we can already confirm our initial intuition about the toy example, by checking that the vector **s1** has a Cost of 2.4, while this is only 0.2 for the optimal solution computed with the algorithm: $\mathbf{s}^* = [1\ 1\ 1\ -1]$ (or its complementary [-1 -1 -1 1]).

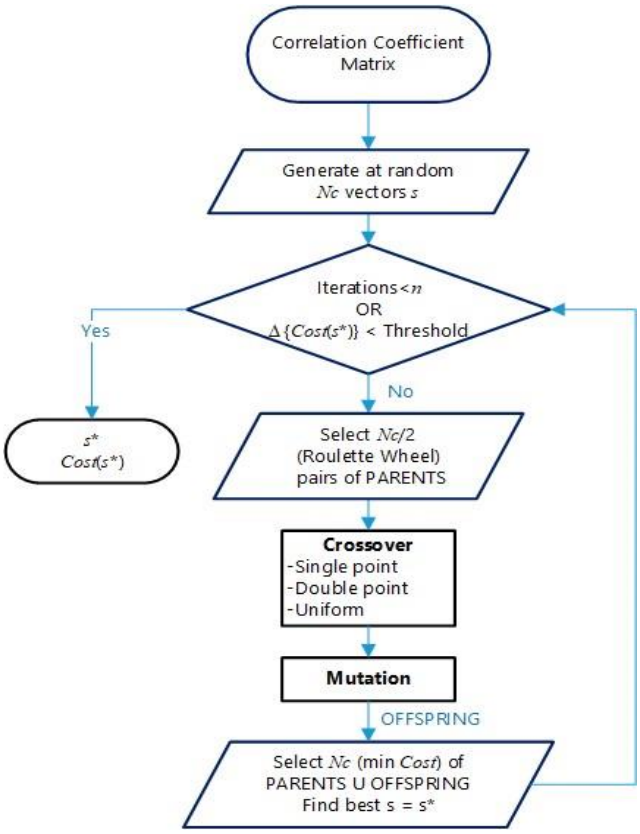


Figure 5.5: Flowchart of the Genetic Algorithm.

5.3.3. Assessment of the genetic-based algorithm for computing polarity

For assessing the results of the algorithm, this section considers the computation of the average image of the 155 ICs (see Section 5.2).

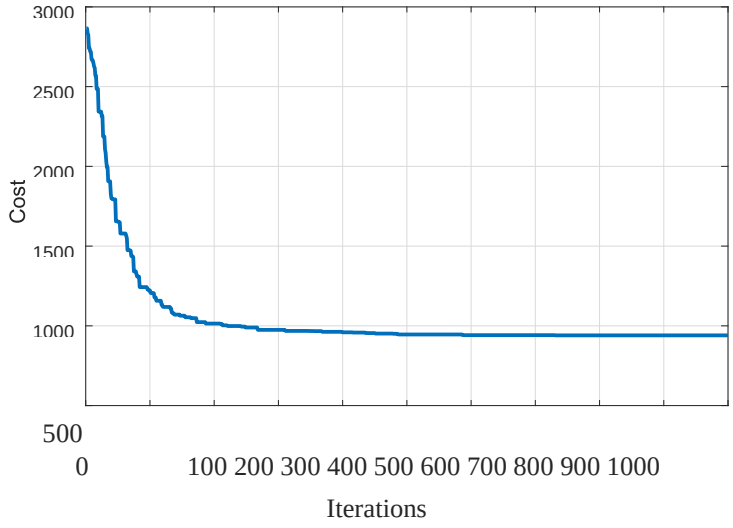


Figure 5.6: Convergence of the genetic algorithm: evolution of the objective Cost function with the number of iterations.

Figure 5.6 plots the evolution of the objective Cost function with the number of iterations. The parameters chosen have been: n = 1000, Nc = 20 and μ = 0.03; the optimal solution $s*$ has a Cost of 941.

The histogram of correct $\mathbf{S}^*_{i,j} = sign(\rho_{i,j})$ and incorrect $\mathbf{S}^*_{i,j} \neq sign(\rho_{i,j})$ for the optimal $\mathbf{s}^*$, for different intervals of absolute correlation coefficients is plotted in Figure 5.7. Not unexpectedly, it shows how the algorithm prioritizes to set correctly the sign for those values with higher absolute value. We can also observe the non-normal distribution of the correlation values [97]. Note that for this figure, the values of the diagonal of the matrix (autocorrelation values) have been removed as they are always correct, regardless of **s**, and therefore do not affect the results.

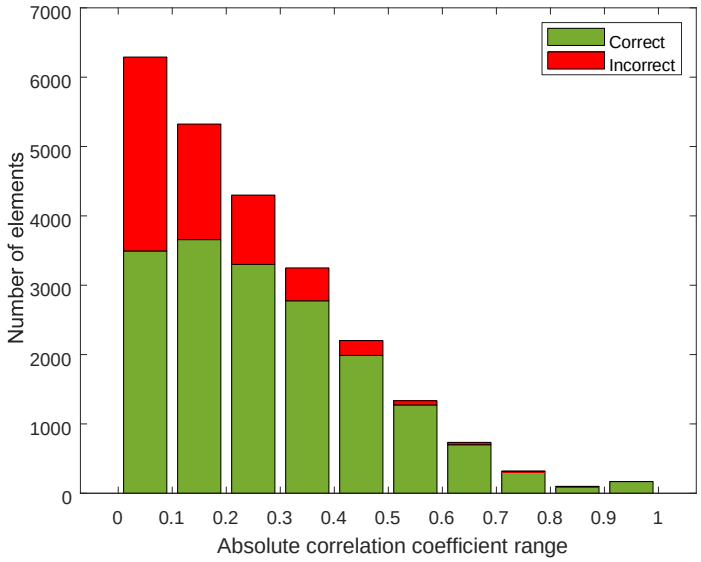


Figure 5.7: Histogram of the corrected/uncorrected values of S^* for the different intervals of absolute correlation coefficients.

Next, these results are compared with those provided by CORRMAP and EEGLAB. Figure 5.8 plots the Cost for the three methods. The values when CORRMAP is used depend on the IC used as reference for the first step. The genetic algorithm provides the best results.

A centroid C, or average image, is then computed using the polarities provided by the three methods. Figure 5.9 compares the *Similarity*, defined as the sum of the absolute correlation coefficients

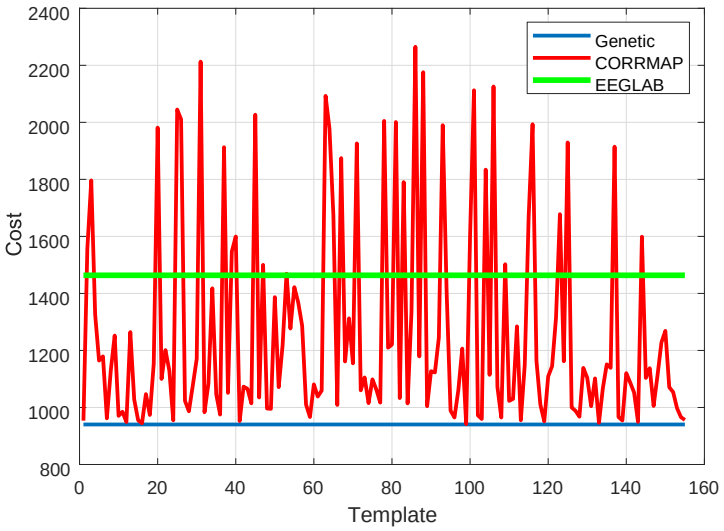


Figure. 5.8: Comparative of the objective Cost.

Comparative of the Cost when using the different methods: genetic algorithm, CORRMAP using each ICs as the template in the first step, and EEGLAB. The genetic algorithm outperforms the rest, providing the lowest value. of the ICs to C, $|\rho_{ic}|$:

$$Similarity = \sum_{i=1}^{155} |\rho_{ic}|. \quad (48)$$

The genetic algorithm outperforms the rest by providing the highest value, indicating a better computed average image.

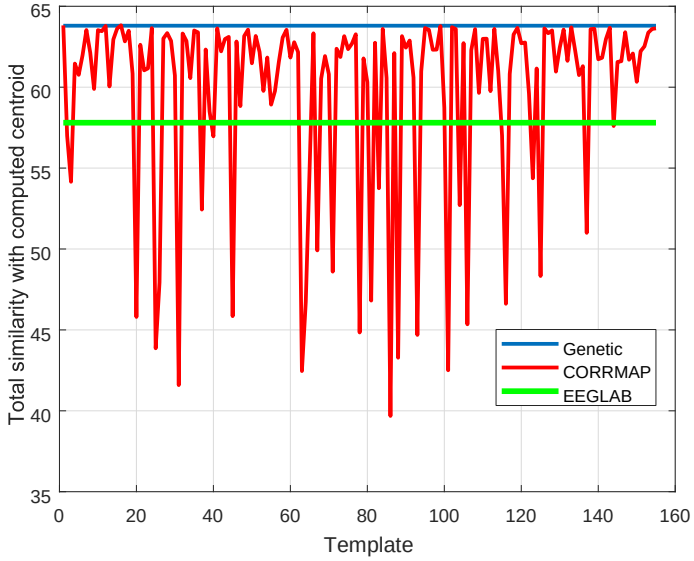
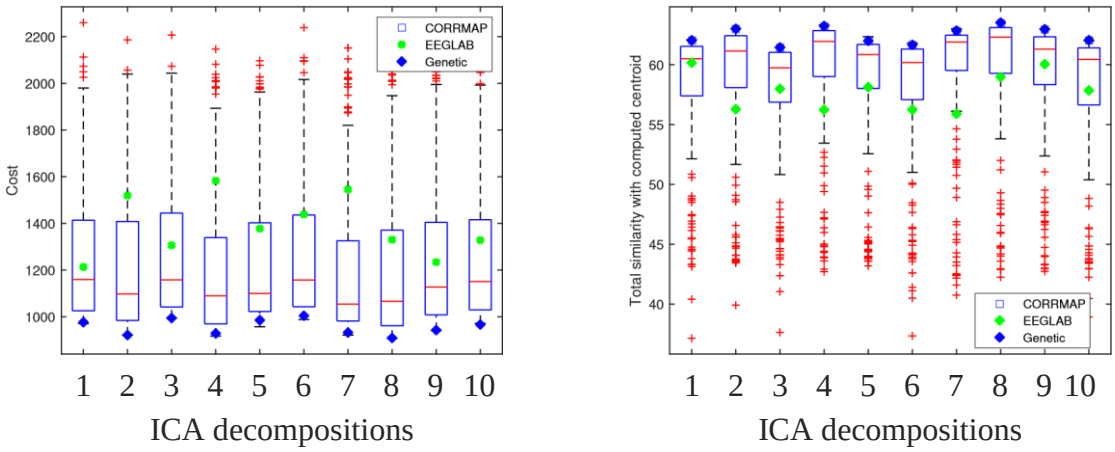


Figure 5.9: Comparative of the Similarity.

Comparative of the Similarity to the computed centroid when using the different methods: genetic algorithm, CORRMAP using each ICs as the template in the first step, and EEGLAB. The genetic algorithm outperforms the rest, providing the highest value.

Finally, to complete the evaluation, the stability of the results are evaluated across several ICA decompositions on the same group of subjects and different groups of subjects. When evaluated for ICA decompositions, Figure 5.10 show the boxplots of the Cost and Total Similarity for the different methods: genetic algorithm, CORRMAP using each ICs as the template in the first step, and EEGLAB. Figure 5.11 repeats the assessment for different groups of subjects. The values provided by EEGLAB for the groups 7, 8 and 10 match those of the genetic algorithm. The results of the assessment confirm that the genetic algorithm provides an optimal bound that is only occasionally reached by CORRMAP and EEGLAB.



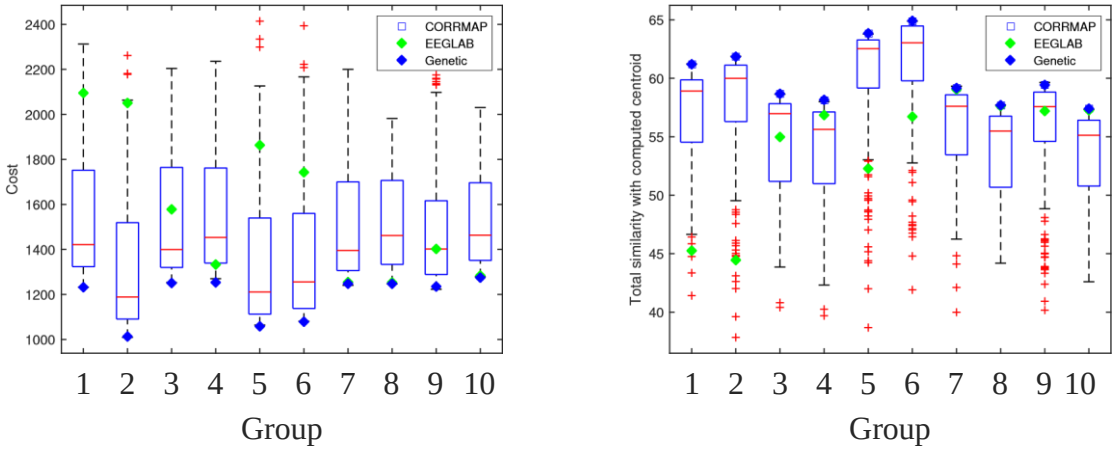
(a) Cost.

(b) Similarity.

Figure 5.10: ICA assessment.

ICA assessment: boxplot of the Cost and similarities with the centroid, computed by different ICA decompositions using the polarities provided by CORRMAP, using different ICs in the first step, the genetic algorithm and EEGLAB.

Group assessment: boxplot of the Cost and similarities with the centroid, computed for different groups of subjects using the polarities provided by CORRMAP, using different ICs in the first step, the genetic algorithm and EEGLAB.



(a) Cost. (b) Similarity.

Figure 5.11: Group assessment.

5.4. Clustering Algorithms for IC topographies

This section reviews the main clustering methods for scalp maps, showing the results for the 155 previously computed ICs. These results are later evaluated across different ICA decompositions and groups of subjects. It is also worth to mention that other clustering algorithms not specifically intended for this purpose, such as DBSCAN [98], were also tried, with conversion from correlations to distances when required, but we decided not to include them here because the obtained results were not relevant enough.

5.4.1. ICLabel

ICLabel classifier is an EEG IC classifier that has shown to perform very well estimating IC classifications as compositional vectors across seven IC categories [99]: 1) brain, activity believed to originate from locally synchronous activity in one (or two well-connected) cortical patches; 2) muscle, high-frequency and broadband components, above 20-30Hz, originated from groups of muscle motor units; 3) eye, which activity originating from the eyes and which can be further subdivided into ICs accounting for activity associated with horizontal eye movements and ICs accounting for blinks and vertical eye movements; 4) heart, they are quite rare and are related to the fact of placing an electrode directly above a superficial vein or artery; 5) Line Noise, concentrated at 50/60Hz and which captures the effects of line current noise emanating from nearby electrical fixtures o poorly grounded EEG amplifiers; 6.) Channel Noise, indicating that some portion of the signal recorded at an electrode channel is already nearly statistically independent of those from other channels; and 7) Others, which

catches those ICs that fit none of the previous types. Thus, ICLabel provides us with a rough initial clustering along with an "estimated" label for the IC components. For this particular case, it results in 3 clusters with: 87 brains, 7 muscles and 15 eyes, leaving 46 as others. The silhouette graph is plotted in Figure 5.12, corresponding to Sh = 0.21, and QI_c of 10.2, 1 and 23 for the clusters Brain, Muscle and Eye, respectively (-1.7 for others), with QI=8 and QI_c=7.6.

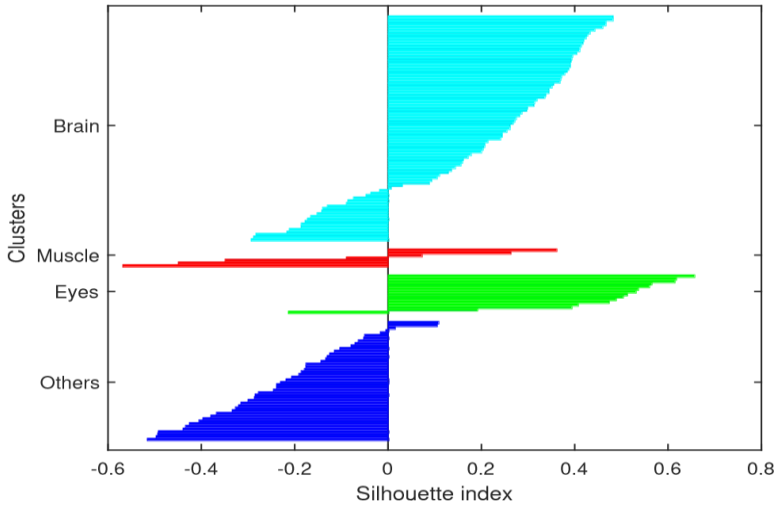
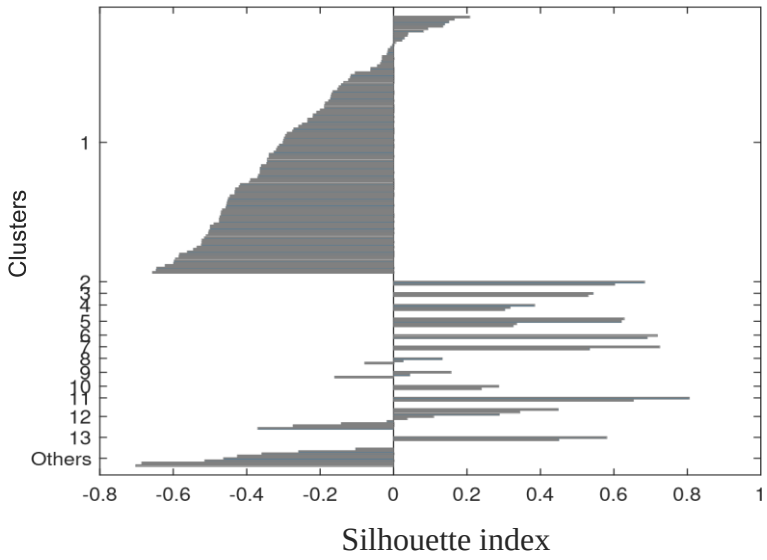


Figure 5.12: Silhouette graph for ICLabel.

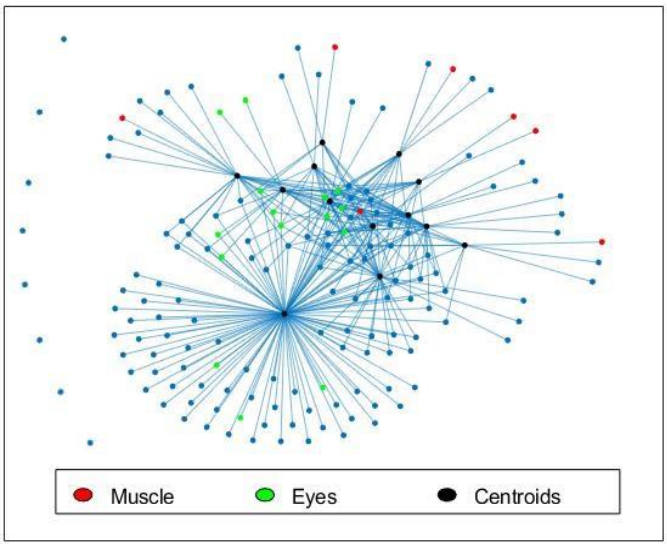
5.4.2. CORRMAP

CORRMAP finds scalp maps that are similar to another selected by the user as a template. It is thus defined as a semi-automatic clustering tool because it does not directly provide the clusters but finds IC topographies that are similar to a user-defined template. The core of the algorithm is a two-step loop. In the first step, the absolute correlation coefficients between a selected IC (template) and the rest of ICs from all datasets are computed. For each dataset, CORRMAP selects up to a number g, chosen by the user from 1 to 3, of ICs with the largest supra threshold absolute correlation with the template. Next, an average cluster map is calculated, after inversion of those ICs showing a negative correlation with the template and root mean square (RMS) normalization of each IC. In the second step, the process is repeated but using the average cluster map obtained in the first step as the new template.

To evaluate CORRMAP clustering performances, we run it 155 times by selecting every IC as the template and looking for similar ICs across subjects (with automatic threshold). A symmetric adjacency matrix is then built setting an edge between the IC selected as template and the ICs found by CORRMAP for such template. The best results are obtained for g = 1. Figure 5.13 shows the silhouette graph of the obtained clustering, with Sh=-0.1, and a graph with the centroid connections. In the latter, clusterization are represented as a graph, with an edge between the centroid m, computed as the average image of the cluster C_m, and any node i provided that $|R_{mi}| \geq min(|R_{mj}|)$ for $\forall j \in C_m$, revealing the relationship between clusters. The figure allows checking at a glance the number of clusters, number of outliers, size of the clusters and separation of the different clusters. Ideally, each element should be connected to a single centroid and each centroid exclusively to the elements of its cluster. Additionally, the figure points out with different colors those ICs labelled by IClabel as eyes (green) or muscle (red), so it is possible to check how these have been clustered. A total of 13 clusters are obtained with QI=15.8, QI_c=11.3 and 8 outliers. It must be noted that there is a big cluster of 111 ICs, which makes even bigger when g increases (125 for g = 2 and 127 for g=3).



(a) Silhouette graph.



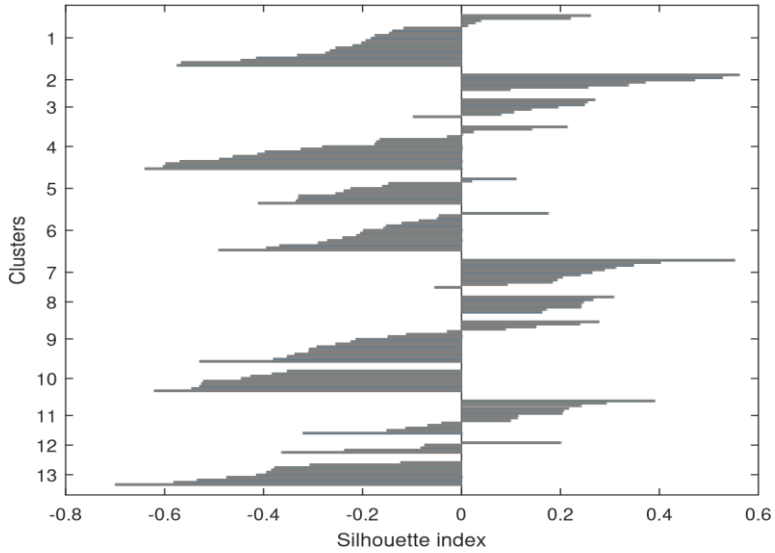
(b)Clusters: centroid connections.

Figure 5.13: Results using CORRMAP.

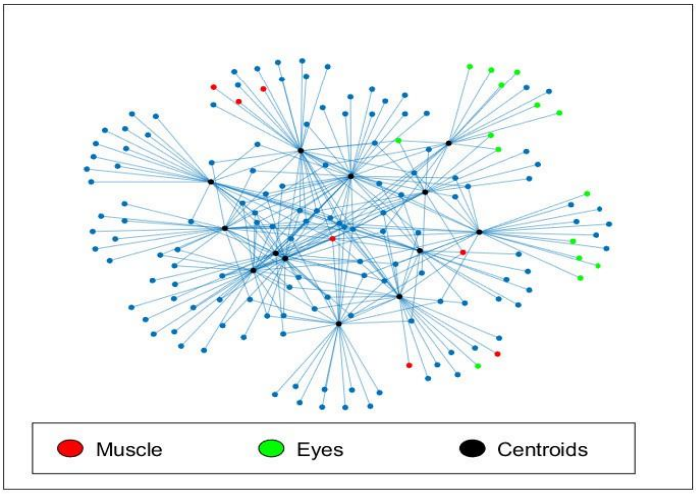
5.4.3. PCA-based built-in EEGLAB clustering algorithms

EEGLAB implements some PCA-based clustering algorithms that allow clustering EEG data according to different characteristics, including the scalp map similarities. The performances provided K-means (Statistics Toolbox), Neural Network and K-meansCluster (Non-Statistic toolbox), henceforth K-meansC, are analyzed here. These methods do not work directly on the inverse ICA weights but on their corresponding topographic map (67x67 matrices).

A grid search for the optimal number of clusters k and number of PCA components p is carried out between 10 and 18, and 3 and 11, respectively, and averaging on 10 realizations for K-means and Neural Network as these algorithms use random initial seeds. The best results, in terms of the cost function QI_c, and the parameters used in each case are described next. For K-means, QI = 14 ± 0.6 and QI_c = 9.8 ± 0.5 are obtained for k = 13, p = 7 and separating as outliers those components to more than 3 standard deviations. For Neural Networks, the best results are obtained for k = 10 and p = 11: QI = 10.7 ± 0.8 and QI_c = 9 ± 1.2. And finally, for K-meansC, QI = 14 and QI_c =9.3 are obtained for k = 14 and p = 6. Figure 5.14 shows the silhouette and the centroid connections for K-means (one of the realizations).



(a) Silhouette graph.



(b) Clusters: centroid connections.

Figure 5.14: Results using Kmeans of EEGLAB.

5.5. Novel Clustering Algorithm

Figure 5.15 shows the flowchart of the novel clustering algorithm proposed in this chapter. A preclusterization based on spectral clustering is followed by a clustering genetic algorithm (CGA).

Spectral clustering [100] is not based on distance but in similarity graphs, which makes it particularly suitable for this case where the absolute correlation coefficient is used as similarity measure. Thus, the proposed clustering algorithm starts by computing a similarity graph *G* as an undirected graph where the edges between two vertices (ICs)

carry a non-negative weight proportional to the similarity measure. More exactly, the adjacency matrix **J** of *G* is computed as: **J** = |**R**| (see Section 5.2.3). Then the symmetric normalized Laplacian is computed [101]:

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{J}\mathbf{D}^{-1/2} \quad (49)$$

where **D** is the degree matrix of *G*.

The next step is computing the first *k* eigenvectors of **L**, corresponding to the *k* smallest eigenvalues. The value *k* is an input parameter of the algorithm and will determine the initial number of pre-clusters. For a first approach to this value, the eigenvalues can be used. From the Laplacian properties, the number of eigenvalues which are (approximately) zero coincides with the number of components (independent clusters) in the graph [102]. The first nonzero eigenvalue is called the eigengap or spectral gap, and informs us about the connectivity of the graph and the number of clusters. As explained later, the final number of clusters may change after applying the second optimization phase. The *k* smallest eigenvectors are then arranged in columns to have a matrix $\mathbf{U} \in \mathrm{R}^{n \times k}$, with n=155 in this case. Using this, the n vectors $y_i \in \mathrm{R}^k$ corresponding to the rows of **U** are clusterized in *k* clusters to get the pre-clustering. These vectors y_i are the coordinates of the original data points in a lower-dimensional space created by the selected eigenvectors. This change of the representation of the data points from R^n to R^k makes clustering easier and is the "key" of the spectral-clustering. Thus, for this step, a simple K-means without outliers' algorithm has been employed. Finally, each original point *i* is assigned to the same cluster that its representation y_i in the reduced dimensional space.

The clustering obtained in the previous phase is then used as the initial seed for the GA implemented in the next phase. This is an elitist algorithm that implements centroid-based encoding. Integer encoding with a vector of N=155 positions is employed [103], allowing that the number of clusters can change during the optimization process. A population of *Nc* children is generated by randomly selecting one of the *N* objects (ICs). For each of these children, one of four possible mutation operators are employed: *merge* and *agglomerative* if the selected IC is an outlier, and *split* and *move*, otherwise. The operators *merge* and *agglomerative*, applied with probability α and $1 - \alpha$, respectively, join the selected IC, which currently is an outlier, to the closest outlier (provided that it exists) to form a new cluster (*merge*) or to the cluster with the closest centroids

(*agglomerative*). The operators *split* and *move*, applied with probability β and $1-\beta$, respectively, assign the selected IC to the group of outliers (*split*) or the cluster with the closest centroid (*move*). The *Nc* resulting clustering's are evaluated and compared with the initial seed. The best solution is chosen as seed for the next iteration.

The process is repeated for up to a maximum of iterations, *MaxIter*, or the results are not improved for a certain number of iterations $\Delta Iter$. This optimization automatically returns the optimum number of clusters provided that *k* in the pre-clustering phase is chosen within a certain range.

For these input parameters: $N = 155$, $k = 13$, $Nc = 20$, $\alpha = 0.5$, $\beta = 0.5$, $MaxIter = 5000$ and $\Delta Iter=250$, Figure 5.16 shows the silhouette graph ($SH = 0.18 \pm 0.01$) and the centroid connections. The number of final clusters is 13.9 ± 0.8 clusters (values from *k* between 9 and 15 converge to around 14), the ICs classified as others is 4.7 ± 1.5, $QI = 26.8 \pm 1.1$ and $QI_c = 24.3 \pm 0.8$. These results clearly outperform those provided by previously analyzed clustering methods. Next section assesses these results across different ICA decompositions and subjects.

5.6. Results

For the assessment of the results, we computed the outputs of the clustering algorithms across 10 ICA decompositions for the same group of subjects and, after that, for 10 different groups of subjects.

Before assessing the results of the clustering algorithm, the reliability of the AMICA decompositions must be analyzed. This reliability is evaluated here by analyzing the results of ICLabel across the 10 ICA decompositions. Figure 5.17(a) graphs the boxplot of the assigned label. The results show great stability (narrow boxes) which is confirmed when performing a χ^2 square test for the observed distributions of the given labels, taking the means as the expected values and six degrees (seven possible labels minus 1) of freedom (see Figure 5.17(b); the *p*-values are above 95% (the green dot indicates the decomposition used in the previous section). Building upon this stability, a new

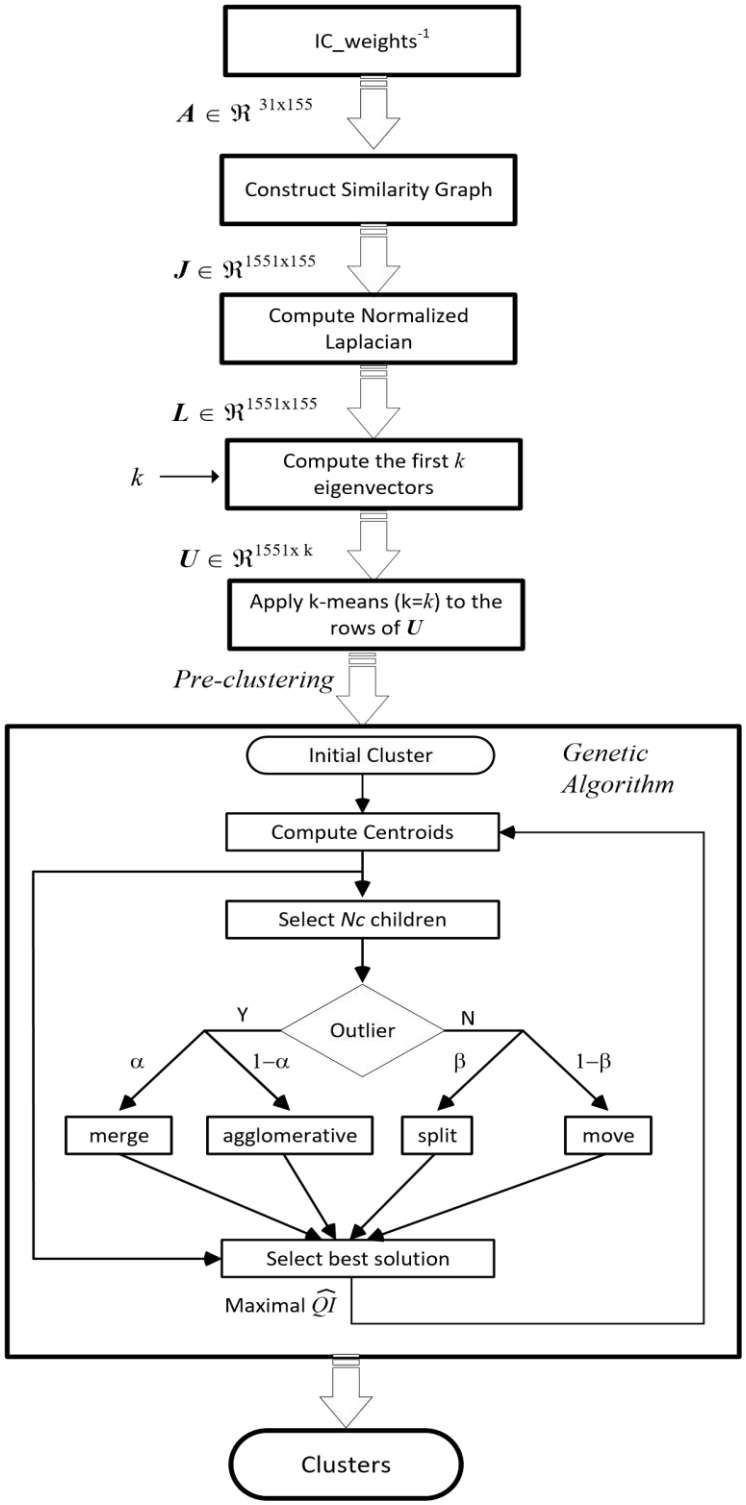
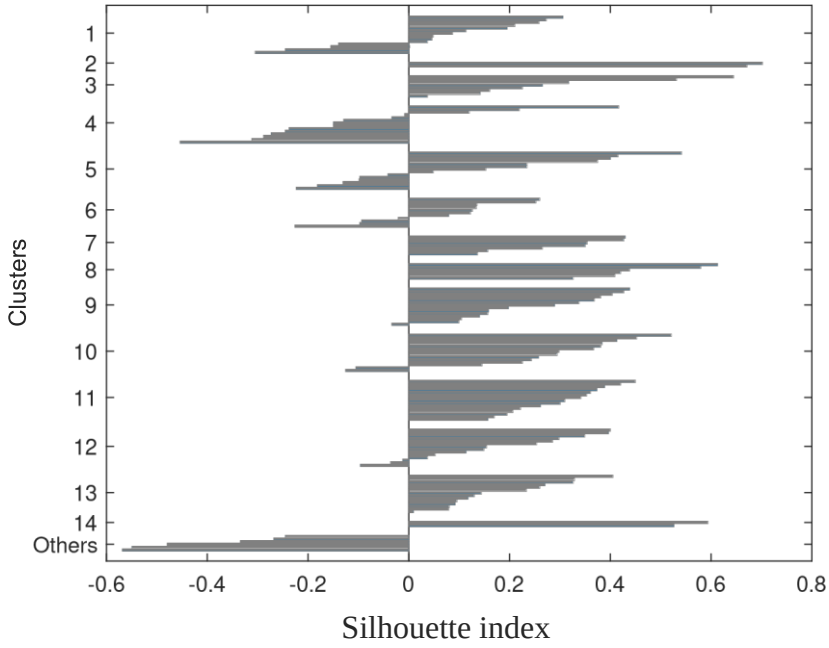
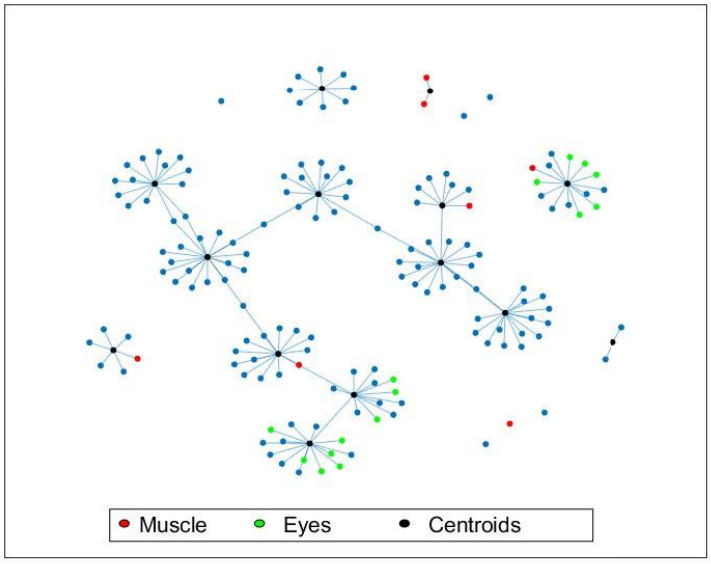


Figure 5.15: Flowchart of the proposed clustering algorithm.



(a) Silhouette graph.



(b) Clusters: centroid connections.

Figure 5.16: Results using the proposed method.

CVI (*RIm*), inspired by *RI* and with values between 0 and 1, is included regarding the ICs labelled as eyes by ICLabel, and computed as follows:

$$RIm = |C_{eye}| \cdot \sum_{C_i \in S_{eye}} \frac{1}{|C_i|} \tag{50}$$

where C_{eye} denotes the eye-cluster generated by ICLABEL, and S_{eye} the set of clusters C_i in the evaluated clustering such that $C_i \cap C_{eye} \neq \emptyset$.

Then, Tables 1 and 2 collect the results across ICA decompositions and groups of subjects, respectively. The relative positions between the clustering methods remain: the proposed method obtains the best results, followed by CORRMAP and K-means and K-meansC, with similary

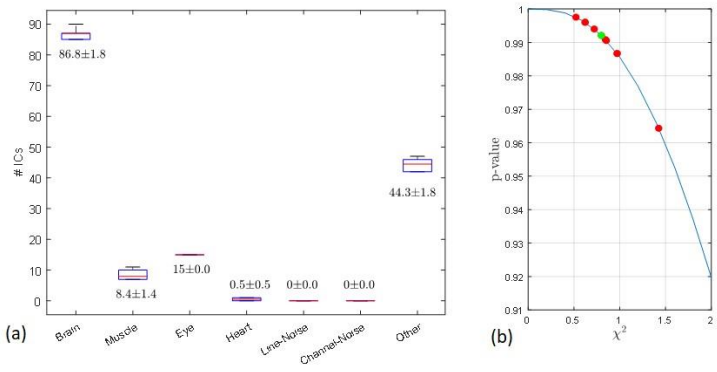


Figure 5.17: AMICA shows great stability

AMICA shows great stability when labels assigned by ICLabel are analyzed. (a) Boxplot of the labels. (b) χ2 scores and the corresponding p-values for the distribution of the assigned labels (the green dot indicates the decomposition used in the previous section).

performances. As expected, the variations of the results are larger for the subject assessment than for the ICA assessment. However, we note that performances of CORRMAP improves when different groups of subjects are tested. This can be explained because the number of clusters is not an input parameter of this algorithm, which allows that it can be freely adjusted for each group of subjects. Even so, the proposed method still outperforms it clearly for all the CVIs.

Table 5.1: Assessment across ICA decompositions

Algorithm	**QI**	***QI***	**Sh** [a]	**DB** [a]	**RIm** [a]	**#clusters**	**#others**
CORRMAP	15.5±1.5	10.9±1.1	-15±3	38±2	13±1	13.4±2.6	6.4±1.1
K-means	13.6 ± 1.1	9.3±1	-12±2	54±2	24±4	12.4±0.5	7.8±5.4
N.Network	10.2±0.8	8.4±1.1	-14±5	57±3	16±2	10	0
K-meansC	13.6±0.4	8.3±0.5	-10±2	53±2	22±3	14	0
Proposed Met.	27.6±0.9	24±0.7	18±2	34±2	46±8	14±0.9	4.9±2.2

[a] ($\cdot 10^{-2}$)

Table 5.2: Assessment across Subjects' groups

	QI	***QI***	***Sh*** [a]	***DB*** [a]	***RIm*** [a]	**#clusters**	**#others**
CORRMAP	20.9±3.9	14.7±3.3	-10±5	42±3	15±5	15.8±2.3	7.3±2.1
K-means	13.7±2	9.4±2	-11±4	53±3	16±5	12.9±0.3	1.6±2.1
N.Network	8.7±2.5	7±3	-15±6	61±3	12±6	9.7±0.6	1.4±3.8
K-meansC	12.8±2.2	7.6±2.3	-13±4	54±3	17±5	14	0
Proposed Met.	28±2.3	25±2.2	17±3	36±3	28±8	13.9 ±1.2	3.3 ±1.1

"(-10^2)

5.7. Conclusions

Clustering of scalp maps is proved to be an effective way to identify relevant source components for ASSR EEGs. The most challenging aspect of this clustering is that traditional euclidean-distance based methods and metrics cannot be directly applied here so they have to be adapted to the use of the absolute correlation coefficient as the similarity measure.

In this chapter we have described a hybrid CGA that dramatically outperforms clustering algorithms provided by EEGLAB; namely, K-means, Neural Network, K-meansCluster and CORRMAP. It consists of a pre-clustering phase based on spectral clustering, which allows a direct adaptation from the pairwise absolute correlation coefficients to similarity graphs, followed by a genetic optimization phase. This optimization phase minimizes a cost function to determine the final partitions, including the number of clusters and the elements which are not assigned to any cluster. This phase implies the computation of centroids that is based on another GA that estimates the polarities of the components.

The performances of the proposed algorithm have been evaluated using specific metrics and assessed across different ICA decompositions and groups of subjects, resulting in the proposed algorithm outperforming significantly the baseline clustering methods. A better clustering of scalp maps should result in a simpler identification of brain-generated processes, so we hope that this work can help researchers working in the field of ASSR EEGs to associate obtained topographic scalp maps with the corresponding populations of interest.

CHAPTER 6: CONCLUSIONS AND FUTURE WORK

6.1. Summary of findings

Bioinformatic is a crucial field at the intersection of biology and computational science. It plays a significant role in analyzing and interpreting biological data.

It's essential to:

- Bioinformatic helps scientists understand vast amounts of biological data. And by analyzing biological datasets to gain insight into how organisms function at the molecular level.
- Bioinformatic plays a pivotal role in drug discovery and development. By analyzing biological data, potential drug targets can be identified, drug efficacy and side effects can be predicted, and drug candidates can be optimized to achieve better therapeutic outcomes.
- The use of bioinformatics has facilitated the implementation of personalized medicine through the genetically analysis of individuals which enables suitable medical treatments to be enacted and therefore improving healthcare services.
- Bioinformatics sheds light on the molecular causes of numerous illnesses – these include heart disease, liver disease, cancer, diabetes and infectious diseases. The variability present across genomic and proteomic data obtained from patients helps establish genetic variants, biomarkers, and disease pathways resulting in more effective disease diagnosis, post-symptoms assessment, and treatment techniques.

6.1.1. The Role of Artificial Intelligence (AI) Algorithms in Bioinformatics

Data Analysis – AI algorithms such as machine learning and deep learning come in handy in bioinformatics data analysis. They further assist in pattern identification, correlation, and other structures lying at the core of large-scale genomic and proteomic

as well as other biological datasets which may have been difficult to identify using traditional statistical techniques.

Predictive Modeling. Bioinformatics has a host of AI algorithms including neural networks that can formulate predictive models surrounding numerous biological activities including protein structure and drug – target interactions.

Feature Selection. Feature selection is a necessity in bioinformatic algorithms that aim at picking out important features in high-dimensional biological data. And AI algorithms can do that – they can choose informative features and significantly affect the model by reducing dimensionality and increasing the model's efficiency.

Drug Discovery and Design: AI algorithms do seem to be a game changer in the drug candidate screening and identification process as they can greatly make the work of searching for and designing new drug candidates faster and easier in every respect. Thus, they are able to estimate the binding affinity of small molecules with target proteins, prepare entirely new drug-like small molecules, and improve lead compounds in terms of their potency and specificity.

Clinical Decision Support: AI algorithms seem to have the potential of assisting physicians in their day to day making of crucial decisions by providing patients demographic with information about the management of the disease which includes its diagnosis, prognosis and even the best therapy to use. They can also synthesize images, sequencing, and clinical records amd other forms of data to enable spine healthcare professionals to practice more tailored and scientific medicine.

As a concluding remark, the combination of bioinformatic and AI is a very strong constructive potential for solving the challenges associated with the biological domains, carrying out fundamental and applied studies and enhancing the quality of life for humanity.

Because healthcare facilities generate a large amount of data, it might be considered a type of big data. This work used a feature selection technique in conjunction with an artificial neural network to sift through the mountains of data. We employ Harris Hawk's optimization approach for feature selection, and for training and testing, we rely on the artificial neural network. Precision, sensitivity, and accuracy are 95.69%, 92.75%, and 92.15%, respectively, for the suggested approach. Compared to MLP,

SVM, RF, and AdaBoost, the suggested strategy improves the accuracy of heart disease diagnoses. Medical data processing will eventually make use of big data systems like Hadoop and Apache Spark.

Patient confidentiality is of the utmost importance in the context of big data's potential uses in healthcare and smart city applications. Our research indicates that blockchain technology is very distributive. It is quite easy to link this device to the IoT. One possibility is to use the blockchain to encrypt private medical records. This problem can be partially or completely resolved with the use of blockchain technology. Using blockchain technology, this work is creating a system to protect patients' privacy. Using machine learning and data mining, every node in the blockchain can initially examine the blockchain data or patient records. Once other hospitals have given their OK, you can add this analysis to the blockchain blocks. This is the suggested approach to the adoption of blockchain technology to ensure the security of patients' personal information during the transmission of their cardiac records and information. Machine learning employs a voting approach to examine data blocks.

During the machine learning stage, the HHO approach enhances classification accuracy. The proposed decentralized approaches are extremely safe when used in the blockchain. In theory, no one should be able to decode the blockchain. Unfortunately, due to the fact that non-distributed systems like blockchain utilize all system components for decoding, they have higher power consumption compared to centralized techniques. Because they employ the blockchain to thwart attacks, the proposed decentralized systems are, in theory, unbreakable. Centralized medical systems, on the other hand, are easy targets for outside interference. One of the suggested approaches is making use of blockchain technology to make decentralized techniques more scalable. No matter the size or shape, the suggested approach will work.

In order to make patient records and data impregnable, the proposed solution uses the blockchain architecture. Like bitcoin, the proposed technique utilizes a lot of RAM due to its distributed nature. Compared to centralized medical systems, the suggested method and blockchain-based decentralized alternatives significantly reduce the time it takes for patients to authenticate. The memory usage of the proposed technique is

higher than the centralized solutions, according to the experiments, because of the blockchain maintenance.

Although blockchain-based authentication takes more time than centralized ones, it is entirely safe because each system's processing time increases when running in distributed computing mode. On the other hand, the data kept in the main system isn't really secure. According to the results, the suggested approach achieves sensitivity of 92.15%, precision of 95.69%, accuracy of 92.75%, and feature selection mechanism of 92.15% when employed simultaneously. Compared to other methods, the suggested one improves upon artificial neural networks, support vector machines, decision trees, random forests, AdaBoost, and Bayesian networks in terms of accuracy in cardiac illness diagnosis. Future medical and health data processing will make use of big data processing platforms like Apache Spark and Hadoop due to the large amounts of data produced by healthcare facilities.

Over the last several decades, researchers have focused extensively on developing methods for automated arrhythmia detection utilizing electrocardiogram (ECG). It is now possible to analyze heart rates from electrocardiograms (ECGs) because of the proliferation and improvement of open-source ECG databases like MIT-BIH and slpdb. Numerous methods exist for classifying heart rates into five categories according to the AAMI standard, which enables one to perceive the waveform of the heartbeat: normal ectopic, extracerebral, ventricular ectopic, fusion, and unknown beat. Using reliable slpdb data, this research set out to create an intelligent medical diagnosis system. The first step of this method is to input data from ECG signals into the program. Next, a deep learning approach known as a convolution neural network [104], There were operations to extract features using a differential evolution optimization technique and classify them with the goal of detection.

The results demonstrate that the proposed method outperforms previous methods for the diagnosis of cardiac arrhythmias, expressed as a percentage. This research considers three main algorithms for identifying coronary heart disease: one based on genetics, one on evolutionary principles, and one on the HHO algorithm.

We compared various methods using the same parameters (the same dataset, the same parametric rate, and different operators) and a variety of criteria to determine how well they performed. For the suggested approaches, the most relevant metrics are sensitivity

(96.04%), accuracy (95.00%), and rate of features (93.94%). The corresponding values for the sensitivity section are 96.04%, 95.54%, and 96.34%. Out of the three feature rates, 82.23%, 82.16%, and 82.41% are present. Not to mention two major publications that introduced an AI-based approach [60] and using deep neural network [61] use the same work's results for the diagnosis of cardiac arrhythmias, regardless of the number of convulsions. In terms of accuracy, they came up with 93.18% and 92.97%, respectively. In comparison to the two sources, the suggested method (a deep convolution neural network based on the HHO algorithm) achieved an accuracy of 95.00%. [61] and [61] and a gain of about 0.86% to 1.22%.

Presenting the work in a scientific context helps overcome one of the biggest challenges facing this work: a dearth of national data, particularly clinical data. Big data, which refers to processing data on a massive scale, also requires powerful technologies.

Liver disease diagnosis framework that uses the Harris Hawks Optimization (HHO) algorithm and the Artificial Neural Network (ANN). This framework was almost 100% accurate and stable in diagnosing liver diseases, beating today's leading methods. We evaluate the performance, compare results and interpret insights to demonstrate its effectiveness and superiority.

We found that 92.5% was the overall accuracy rate of the developed framework. Specifically, precision and recall scores were 89% and 94% respectively. This demonstrates the advanced diagnostic ability of the ANN--it can indeed accurately identify liver disease conditions. When considering our comparative analysis, the new proposed framework outshines the existing methods consistently, surpassing them in accuracy, sensitivity and specificity. Given these new results, the optimization approach that is HHO-based and combined with ANN offers much promise for diagnosing liver diseases.

The current work involved deposition of the VGG16 architecture as an effective Artificial Neural Network (ANN) for feature extraction from preprocessed iris images. The amalgamation of ANN with the recently developed Harris Hawks Optimization (HHO) algorithm was tested using a benchmark iris dataset to accomplish the purpose of iris detection. Results depict that the HHO-ANN model surpassed the traditional VGG16 model in accuracy, preciseness, recall and F1 score. The performance differences were also found statistically significant. Moreover, the HHO-ANN model

showed competitive computational efficiency against other methods including recent deep learning techniques. These outcomes reveal the promise and efficacy of the HHO-ANN model with VGG16 feature extraction for refining state-of-the-art iris detection systems and making them a suitable selection for deploying in biometric identification domain.

This document successfully proved that the proposed integrated HHO-ANN model with VGG16 feature extraction when employed for iris detection, performs quite well. However, the future works may include employing different deep learning architectures for feature extraction; their comparison with the HHO-ANN model can unravel the optimal deep learning architecture for the iris recognition. The work of transfer learning can be employed in which the pre-trained models are adapted and fine-tuned on the iris datasets. This will ameliorate the performance and efficiency of the iris detection system.

As it turns out, clustering scalp maps is a great technique to group ICs for EEGs taken when patients are at rest. Previously proposed clustering methods like CORRMAP or the built-in algorithms of EEGLAB are outperformed by Spectral-clustering, which exhibits the best combined results in terms of Quality Index and represented elements. The advantage of Spectral clustering can be explained because it does not work with distance measures but with similarity graphs, which allows a more direct adaptation from pairwise absolute correlation coefficients.

The full process of grouping scalp topographic EEG maps was detailed in this article. The most challenging aspect of this clustering is adjusting to the use of the absolute correlation coefficient as a measure of similarity rather than the more traditional metrics and techniques based on euclidean distance. This leads us to suggest a new hybrid clustering approach that combines spectral-clustering with genetic optimization. Using tailored measures, we assessed this algorithm's efficacy and found that it much surpasses the baseline clustering techniques. Since better grouping of these maps should make it simpler to identify brain-generated processes, we anticipate that our work can help researchers in the field of ASSR EEGs associate the topographic scalp maps with the appropriate populations.

6.2. Future work

There are opportunities for future research to explore different combination methods, incorporate additional clinical variables, and enhance the accuracy and reliability of heart disease diagnosis models. These advancements would benefit patients by improving medical decision-making and patient outcomes.

The future works include employing different deep learning architectures for feature extraction; their comparison with the HHO-ANN model can unravel the optimal deep learning architecture for the iris recognition. One approach is to use transfer learning, which involves fine-tuning pre-trained models using the iris datasets. Because of this, the iris detecting system will work better and faster. To test the HHO-ANN model's generalizability, future work should include expanding the work to bigger and more difficult iris databases. Deploying the HHO-ANN model on platforms with limited resources, including embedded systems or mobile devices, requires investigation into their practicality.

REFERENCES

[1] H. Al-Safi, J. Munilla, and J. Rahebi, "Harris Hawks Optimization (HHO) Algorithm based on Artificial Neural Network for Heart Disease Diagnosis," in *2021 IEEE International Conference on Mobile Networks and Wireless Communications (ICMNWC)*, 2021, pp. 1–5.

[2] H. Al-Safi, J. Munilla, and J. Rahebi, "Patient privacy in smart cities by blockchain technology and feature selection with Harris Hawks Optimization (HHO) algorithm and machine learning," *Multimed. Tools Appl.*, pp. 1–25, 2022.

[3] H. Alsafi, J. Munilla, and J. Rahebi, "An Approach for Cardiac Coronary Detection of Heart Signal Based on Harris Hawks Optimization and Multichannel Deep Convolutional Learning," *Comput. Intell. Neurosci.*, vol. 2022, 2022.

[4] H. Alsafi, H. Alsalihi, and J. Munilla, "Use Harris Hawks Optimization (HHO) Algorithm based on Artificial Neural Network for liver disease diagnosis ," *Proc. Int. Conf. Intell. Syst. New Appl.*, vol. 2, no. SE-Proceedings Paper, pp. 1–10, Apr. 2024.

[5] H. E. S. Al-Safi and J. Munilla, "A Neural Network-Based Harris Hawks Optimization Algorithm for Iris Detection." Still under process.

[6] J. Munilla, H. E. S. Al-Safi, A. Ortiz, and J. L. Luque, "Hybrid Genetic Algorithm for Clustering IC Topographies of EEGs," *Brain Topogr.*, vol. 36, no. 3, pp. 338–349, 2023.

[7] J. Munilla, A. Ortiz, H. E. S. AlSafi, and J. L. Luque, "Enhanced Computation of the EEG-IC Polarities using a Genetic Algorithm," in *2022 12th International Congress on Advanced Applied Informatics (IIAI-AAI)*, 2022, pp. 319–324.

[8] S. Roy and K. I. Shoghi, "Computer-Aided Tumor Segmentation from T2-Weighted MR Images of Patient-Derived Tumor Xenografts," in *International Conference on Image Analysis and Recognition*, 2019, pp. 159–171.

[9] S. Agrawal and J. Agrawal, "Neural network techniques for cancer prediction: A survey," in *Procedia Computer Science*, 2015, vol. 60, no. 1, pp. 769–774.

[10] S. Roy *et al.*, "Co-clinical FDG-PET radiomic signature in predicting response to neoadjuvant chemotherapy in triple-negative breast cancer," *Eur. J. Nucl. Med. Mol. Imaging*, vol. 49, no. 2, pp. 550–562, 2022.

[11] S. Mishra, K. Shaw, and D. Mishra, "A New Meta-heuristic Bat Inspired Classification Approach for Microarray Data," *Procedia Technol.*, vol. 4, pp. 802–806, 2012.

[12] Y. Dagli, S. Choksi, and S. Roy, "Prediction of two year survival among patients of non-small cell lung cancer," in *Computer Aided Intervention and Diagnostics in Clinical and Medical Images*, Springer, 2019, pp. 169–177.

[13] M. A. Karaolis, J. A. Moutiris, D. Hadjipanayi, and C. S. Pattichis, "Assessment of the risk factors of coronary heart events based on data mining with decision trees," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 3, pp. 559–566, 2010.

[14] M. Niroee, "Simulation of a hybrid model using genetic algorithms and artificial neural networks for the differentiation of benign and malignant patterns in breast cancer and mammography," *Iran J Med Phys2006*, vol. 3, pp. 67–80.

[15] J. Ternacle, N. Côté, L. Krapf, A. Nguyen, M.-A. Clavel, and P. Pibarot, "Chronic kidney disease and the pathophysiology of valvular heart disease," *Can. J. Cardiol.*, vol. 35, no. 9, pp. 1195–1207, 2019.

[16] A. A. House *et al.*, "Heart failure in chronic kidney disease: conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference," *Kidney Int.*, vol. 95, no. 6, pp. 1304–1317, 2019.

[17] T. Nguyen and Z. A. Wang, "Cardiovascular screening and early detection of heart disease in adults with chronic kidney disease," *J. Nurse Pract.*, vol. 15, no. 1, pp. 34–40, 2019.

[18] R. Sameni and G. D. Clifford, "A review of fetal ECG signal processing; issues and promising directions," *Open Pacing. Electrophysiol. Ther. J.*, vol. 3, p. 4, 2010.

[19] B. Latré, "Reliable and energy efficient network protocols for wireless body area networks." Ghent University, 2008.

[20] P. Bhavsar, I. Safro, N. Bouaynaya, R. Polikar, and D. Dera, "Machine Learning

in Transportation Data Analytics," in *Data Analytics for Intelligent Transportation Systems*, Elsevier, 2017, pp. 283–307.

[21] W.-Y. Deng, Y.-S. Ong, P. S. Tan, and Q.-H. Zheng, "Online sequential reduced kernel extreme learning machine," *Neurocomputing*, vol. 174, pp. 72–84, 2016.

[22] A. F. A. Iswisi, O. Karan, and J. Rahebi, "Diagnosis of Multiple Sclerosis Disease in Brain Magnetic Resonance Imaging Based on the Harris Hawks Optimization Algorithm," *Biomed Res. Int.*, vol. 2021, 2021.

[23] H. Alabool, D. Al- Arabiat, L. Abualigah, and A. A. Heidari, "Harris hawks optimization: a comprehensive review of recent variants and applications," *Neural Comput. Appl.*, vol. 33, Aug. 2021.

[24] A. Dhiman, K. Gupta, and D. K. Sharma, "Chapter 1 - An introduction to deep learning applications in biometric recognition," in *Hybrid Computational Intelligence for Pattern Analysis*, V. Piuri, S. Raj, A. Genovese, and R. B. T.-T. in D. L. M. Srivastava, Eds. Academic Press, 2021, pp. 1–36.

[25] R. Chakraborty, C. R. Rao, and P. K. Sen, "Introduction: Wither Bioinformatics in Human Health and Heredity," vol. 28, R. Chakraborty, C. R. Rao, and P. B. T.-H. of S. Sen, Eds. Elsevier, 2012, pp. 1–10.

[26] G. Muhammad and M. S. Hossain, "A Deep-Learning-Based Edge-Centric COVID-19-Like Pandemic Screening and Diagnosis System within a B5G Framework Using Blockchain," *IEEE Netw.*, vol. 35, no. 2, pp. 74–81, 2021.

[27] X. Liu, P. Zhou, T. Qiu, and D. O. Wu, "Blockchain-enabled contextual online learning under local differential privacy for coronary heart disease diagnosis in mobile edge computing," *IEEE J. Biomed. Heal. Informatics*, vol. 24, no. 8, pp. 2177–2188, 2020.

[28] N. Jia, S. Fu, and M. Xu, "Privacy-Preserving Nonlinear SVM Classifier Training Based on Blockchain," in *International Symposium on Security and Privacy in Social Networks and Big Data*, 2020, pp. 278–288.

[29] Z. Ma, J. Ma, Y. Miao, and X. Liu, "Privacy-preserving and high-accurate outsourced disease predictor on random forest," *Inf. Sci. (Ny).*, vol. 496, pp. 225–241, 2019.

[30] D. Jayaraj and S. Sathiamoorthy, “Random forest based classification model for lung cancer prediction on computer tomography images,” in *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2019, pp. 100–104.

[31] N. Mohan, V. Jain, and G. Agrawal, “Heart Disease Prediction Using Supervised Machine Learning Algorithms,” in *2021 5th International Conference on Information Systems and Computer Networks, ISCON 2021*, 2021, vol. 136, p. 104672.

[32] A. Mert, N. Kılıç, and A. Akan, “Evaluation of bagging ensemble method with time-domain feature extraction for diagnosing of arrhythmia beats,” *Neural Comput. Appl.*, vol. 24, no. 2, pp. 317–326, 2014.

[33] Y.-P. Huang, C.-Y. Huang, and S.-I. Liu, “Hybrid intelligent methods for arrhythmia detection and geriatric depression diagnosis,” *Appl. Soft Comput.*, vol. 14, pp. 38–46, 2014.

[34] M. Mitra and R. K. Samanta, “Cardiac arrhythmia classification using neural networks with selected features,” *Procedia Technol.*, vol. 10, pp. 76–84, 2013.

[35] S. K. Saini and R. Gupta, “Artificial intelligence methods for analysis of electrocardiogram signals for cardiac abnormalities: State-of-the-art and future challenges,” *Artif. Intell. Rev.*, pp. 1–47, 2021.

[36] B. Zhu, Y. Ding, and K. Hao, “Multiclass maximum margin clustering via immune evolutionary algorithm for automatic diagnosis of electrocardiogram arrhythmias,” *Appl. Math. Comput.*, vol. 227, pp. 428–436, 2014.

[37] R. J. Martis *et al.*, “Computer aided diagnosis of atrial arrhythmia using dimensionality reduction methods on transform domain representation,” *Biomed. Signal Process. Control*, vol. 13, no. 1, pp. 295–305, 2014.

[38] C. Qu *et al.*, “Improving feature selection performance for classification of gene expression data using Harris Hawks optimizer with variable neighborhood learning,” *Brief. Bioinform.*, vol. 22, no. 5, p. bbab097, 2021.

[39] M. Abdel-Basset, W. Ding, and D. El-Shahat, “A hybrid Harris Hawks optimization algorithm with simulated annealing for feature selection,” *Artif.*

Intell. Rev., vol. 54, pp. 593–637, 2021.

[40] S. Afreen, A. K. Bhurjee, and R. M. Aziz, "Gene selection with Game Shapley Harris hawks optimizer for cancer classification," *Chemom. Intell. Lab. Syst.*, vol. 242, p. 104989, 2023.

[41] J. P. Kelwade and S. S. Salankar, "Prediction of cardiac arrhythmia using artificial neural network," *Int. J. Comput. Appl.*, vol. 115, no. 20, 2015.

[42] A. Kondababu, V. Siddhartha, B. H. K. B. Kumar, and B. Penumutchi, "A comparative study on machine learning based heart disease prediction," *Mater. Today Proc.*, 2021.

[43] A. A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. Mafarja, and H. Chen, "Harris hawks optimization: Algorithm and applications," *Futur. Gener. Comput. Syst.*, vol. 97, pp. 849–872, 2019.

[44] S. M. Boubakar Khalifa Albargathe, E. Kamberli, F. Kandemirli, and J. Rahebi, "Blood vessel segmentation and extraction using H-minima method based on image processing techniques," *Multimed. Tools Appl.*, vol. 80, no. 2, pp. 2565–2582, 2021.

[45] F. A. Alsarori, H. Kaya, J. Rahebi, D. E. Popescu, and D. J. Hemanth, "Cancer cell detection through histological nuclei images applying the hybrid combination of artificial bee colony and particle swarm optimization algorithms," *Int. J. Comput. Intell. Syst.*, vol. 13, no. 1, pp. 1507–1516, 2020.

[46] I. A. Masoud Abdulhamid, A. Sahiner, and J. Rahebi, "New Auxiliary Function with Properties in Nonsmooth Global Optimization for Melanoma Skin Cancer Segmentation," *Biomed Res. Int.*, vol. 2020, 2020.

[47] A. S. Abdullah, J. Rahebi, Y. E. Özok, and M. Aljanabi, "A new and effective method for human retina optic disc segmentation with fuzzy clustering method based on active contour model," *Med. Biol. Eng. Comput.*, vol. 58, no. 1, pp. 25–37, 2020.

[48] A. Tandon, A. Dhir, N. Islam, and M. Mäntymäki, "Blockchain in healthcare: A systematic literature review, synthesizing framework and future research agenda," *Comput. Ind.*, vol. 122, 2020.

[49] K. Azbeg, O. Ouchetto, S. J. Andaloussi, and L. Fetjah, "A Taxonomic Review of the Use of IoT and Blockchain in Healthcare Applications," *IRBM*, 2021.

[50] H. M. Hussien, S. M. Yasin, N. I. Udzir, M. I. H. Ninggal, and S. Salman, "Blockchain technology in the healthcare industry: Trends and opportunities," *J. Ind. Inf. Integr.*, vol. 22, p. 100217, 2021.

[51] G. N. Nguyen, N. H. Le Viet, M. Elhoseny, K. Shankar, B. B. Gupta, and A. A. A. El-Latif, "Secure blockchain enabled Cyber–physical systems in healthcare using deep belief network with ResNet model," *J. Parallel Distrib. Comput.*, vol. 153, pp. 150–160, 2021.

[52] R. W. Ahmad, K. Salah, R. Jayaraman, I. Yaqoob, S. Ellahham, and M. Omar, "The role of blockchain technology in telehealth and telemedicine," *Int. J. Med. Inform.*, vol. 148, p. 104399, 2021.

[53] N. Iqbal, F. Jamil, S. Ahmad, and D. Kim, "A Novel Blockchain-Based Integrity and Reliable Veterinary Clinic Information Management System Using Predictive Analytics for Provisioning of Quality Health Services," *IEEE Access*, vol. 9, pp. 8069–8098, 2021.

[54] P. G. Shynu, V. G. Menon, R. L. Kumar, S. Kadry, and Y. Nam, "Blockchain-Based Secure Healthcare Application for Diabetic-Cardio Disease Prediction in Fog Computing," *IEEE Access*, vol. 9, pp. 45706–45720, 2021.

[55] B. A. Y. Alqaralleh, T. Vaiyapuri, V. S. Parvathy, D. Gupta, A. Khanna, and K. Shankar, "Blockchain-assisted secure image transmission and diagnosis model on Internet of Medical Things Environment," *Pers. Ubiquitous Comput.*, 2021.

[56] T. Veeramakali, R. Siva, B. Sivakumar, P. C. Senthil Mahesh, and N. Krishnaraj, "An intelligent internet of things-based secure healthcare framework using blockchain technology with an optimal deep learning model," *J. Supercomput.*, 2021.

[57] A. H. Mohsin *et al.*, "PSO–Blockchain-based image steganography: towards a new method to secure updating and sharing COVID-19 data in decentralised hospitals intelligence architecture," *Multimed. Tools Appl.*, vol. 80, no. 9, pp. 14137–14161, 2021.

[58] H. Kriplani, B. Patel, and S. Roy, “Prediction of chronic kidney diseases using deep artificial neural network technique,” in *Computer aided intervention and diagnostics in clinical and medical images*, Springer, 2019, pp. 179–187.

[59] S. Molaei, N. Ghorbani, F. Dashtiahangar, M. Peivandi, Y. Pourasad, and M. Esmaeili, “FDCNet: Presentation of the Fuzzy CNN and Fractal Feature Extraction for Detection and Classification of Tumors,” *Comput. Intell. Neurosci.*, vol. 2022, 2022.

[60] U. R. Acharya *et al.*, “Automated identification of shockable and non-shockable life-threatening ventricular arrhythmias using convolutional neural network,” *Futur. Gener. Comput. Syst.*, vol. 79, pp. 952–959, 2018.

[61] Q. Yao, R. Wang, X. Fan, J. Liu, and Y. Li, “Multi-class Arrhythmia detection from 12-lead varied-length ECG using Attention-based Time-Incremental Convolutional Neural Network,” *Inf. Fusion*, vol. 53, pp. 174–182, 2020.

[62] J. Sun, S. Zhao, Y. Yu, X. Wang, and L. Zhou, “Iris recognition based on local circular Gabor filters and multi-scale convolution feature fusion network,” *Multimed. Tools Appl.*, vol. 81, Sep. 2022.

[63] Y. Liu, A. Haridevan, H. Schofield, and J. Shan, *Ghost-DeblurGAN and Its Application to Fiducial Marker System*. 2021.

[64] M. V. R. Manimala, C. Dhanunjaya Naidu, and M. N. Giri Prasad, “Sparse MR Image Reconstruction Considering Rician Noise Models: A CNN Approach,” *Wirel. Pers. Commun.*, vol. 116, no. 1, pp. 491–511, 2021.

[65] S. Mukherjee, “The Annotated ResNet-50. Towards Data Science. 2022. Available online: https://towardsdatascience.com/the-annotated-resnet-50-a6c536034758 (accessed on 12 June 2022).”

[66] A. J. Bell and T. J. Sejnowski, “An information-maximization approach to blind separation and blind deconvolution.,” *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, Nov. 1995.

[67] X. Yin, T. Shu, and Q. Huang, “Semi-supervised fuzzy clustering with metric learning and entropy regularization,” *Knowledge-Based Syst.*, vol. 35, pp. 304–311, 2012.

[68] N. Piroonsup and S. Sinthupinyo, “Analysis of training data using clustering to improve semi-supervised self-training,” *Knowledge-Based Syst.*, vol. 143, pp. 65–80, 2018.

[69] E. D. Farahani, J. Wouters, and A. van Wieringen, “Brain mapping of auditory steady-state responses: A broad view of cortical and subcortical sources.,” *Hum. Brain Mapp.*, vol. 42, no. 3, pp. 780–796, Feb. 2021.

[70] E. Hwang, H.-B. Han, J. Y. Kim, and J. H. Choi, “High-density EEG of auditory steady-state responses during stimulation of basal forebrain parvalbumin neurons,” *Sci. Data*, vol. 7, no. 1, p. 288, 2020.

[71] D. Koshiyama *et al.*, “Source decomposition of the frontocentral auditory steady-state gamma band response in schizophrenia patients and healthy subjects.,” *Psychiatry Clin. Neurosci.*, vol. 75, no. 5, pp. 172–179, May 2021.

[72] O. H. Jefsen, Y. Shtyrov, K. M. Larsen, and M. J. Dietz, “The 40-Hz auditory steady-state response in bipolar disorder: A meta-analysis.,” *Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol.*, vol. 141, pp. 53–61, Sep. 2022.

[73] N. J. Gallego-Molina, A. Ortiz, F. J. Martínez-Murcia, M. A. Formoso, and A. Giménez, “Complex network modeling of EEG band coupling in dyslexia: An exploratory analysis of auditory processing and diagnosis,” *Knowledge-Based Syst.*, vol. 240, p. 108098, 2022.

[74] C.-T. Lin, S.-A. Chen, T.-T. Chiu, H.-Z. Lin, and L.-W. Ko, “Spatial and temporal EEG dynamics of dual-task driving performance.,” *J. Neuroeng. Rehabil.*, vol. 8, p. 11, Feb. 2011.

[75] “Makoto’s accessed: 2023-01-19. M. Miyakoshi, “No Title,” *preprocessing pipeline,” https://sccn.ucsd.edu/wiki/Makoto’s preprocessing pipeline,* 2023.

[76] F. C. Viola, J. Thorne, B. Edmonds, T. Schneider, T. Eichele, and S. Debener, “Semi-automatic identification of independent components representing EEG artifact.,” *Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol.*, vol. 120, no. 5, pp. 868–877, May 2009.

[77] F. Artoni, D. Menicucci, A. Delorme, S. Makeig, and S. Micera, “RELICA: A method for estimating the reliability of independent components,” *Neuroimage*,

vol. 103, pp. 391–400, 2014.

[78] F. Artoni, A. Delorme, and S. Makeig, “Applying dimension reduction to EEG data by Principal Component Analysis reduces the quality of its subsequent Independent Component decomposition.,” *Neuroimage*, vol. 175, pp. 176–187, Jul. 2018.

[79] F. G. A. de Meneses, A. S. Teles, M. Nunes, D. da Silva Farias, and S. Teixeira, “Neural Networks to Recognize Patterns in Topographic Images of Cortical Electrical Activity of Patients with Neurological Diseases,” *Brain Topogr.*, vol. 35, no. 4, pp. 464–480, 2022.

[80] A. Delorme and S. Makeig, “EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis.,” *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, Mar. 2004.

[81] A. Ortiz, F. J. Martinez-Murcia, J. L. Luque, A. Giménez, R. Morales-Ortega, and J. Ortega, “Dyslexia Diagnosis by EEG Temporal and Spectral Descriptors: An Anomaly Detection Approach.,” *Int. J. Neural Syst.*, vol. 30, no. 7, p. 2050029, Jul. 2020.

[82] M. Klug and K. Gramann, “Identifying key factors for improving ICA-based decomposition of EEG data in mobile and stationary experiments.,” *Eur. J. Neurosci.*, vol. 54, no. 12, pp. 8406–8420, Dec. 2021.

[83] S. D. Muthukumaraswamy, “High-frequency brain activity and muscle artifacts in MEG/EEG: a review and recommendations.,” *Front. Hum. Neurosci.*, vol. 7, p. 138, 2013.

[84] P. Mitra and H. Bokil, *Observed Brain Dynamics*. Oxford University Press, 2007.

[85] S. Haufe *et al.*, “On the interpretation of weight vectors of linear models in multivariate neuroimaging.,” *Neuroimage*, vol. 87, pp. 96–110, Feb. 2014.

[86] T.-W. Lee, M. Girolami, A. J. Bell, and T. J. Sejnowski, “A unifying information-theoretic framework for independent component analysis,” *Comput. Math. with Appl.*, vol. 39, no. 11, pp. 1–21, 2000.

[87] A. Delorme, J. Palmer, J. Onton, R. Oostenveld, and S. Makeig, “Independent

EEG sources are dipolar.," *PLoS One*, vol. 7, no. 2, p. e30135, 2012.

[88] J. Onton, M. Westerfield, J. Townsend, and S. Makeig, "Imaging human EEG dynamics using independent component analysis.," *Neurosci. Biobehav. Rev.*, vol. 30, no. 6, pp. 808–822, 2006.

[89] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu, and S. Wu, "Understanding and Enhancement of Internal Clustering Validation Measures," *IEEE Trans. Cybern.*, vol. 43, no. 3, pp. 982–994, 2013.

[90] W. Sheng, S. Swift, L. Zhang, and X. Liu, "A weighted sum validity function for clustering with a hybrid niching genetic algorithm," *IEEE Trans. Syst. Man, Cybern. Part B*, vol. 35, no. 6, pp. 1156–1167, 2005.

[91] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Stat.*, vol. 3, no. 1, pp. 1–27, Jan. 1974.

[92] M. S. Rahim, K. A. Nguyen, R. A. Stewart, T. Ahmed, D. Giurco, and M. Blumenstein, "A clustering solution for analyzing residential water consumption patterns," *Knowledge-Based Syst.*, vol. 233, p. 107522, Dec. 2021.

[93] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[94] A. J. Gates and Y. Y. Ahn, "The impact of random models on clustering similarity," *J. Mach. Learn. Res.*, vol. 18, pp. 1–28, 2017.

[95] L. Pion-Tonachini, K. Kreutz-Delgado, and S. Makeig, "ICLabel: An automated electroencephalographic independent component classifier, dataset, and website.," *Neuroimage*, vol. 198, pp. 181–197, Sep. 2019.

[96] R. Tinós, L. Zhao, F. Chicano, and D. Whitley, "NK Hybrid Genetic Algorithm for Clustering," *IEEE Trans. Evol. Comput.*, vol. 22, no. 5, pp. 748–761, 2018.

[97] R. A. Fisher, "Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population," *Biometrika*, vol. 10, no. 4. p. 507, 1915.

[98] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "Why and How You Should (Still) Use DBSCAN," *ACM Trans. Database Syst.*, vol. 42, no. 3, pp.

1–21, 2017.

[99] L. Pion-Tonachini, K. Kreutz-Delgado, and S. Makeig, "The ICLabel dataset of electroencephalographic (EEG) independent component (IC) features," *Data Br.*, vol. 25, p. 104101, 2019.

[100] A. Ng, M. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*, 2001, vol. 14.

[101] S. Butler and F. Chung, "Spectral Graph Theory," *Handb. Linear Algebr. Second Ed.*, 2013.

[102] U. von Luxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[103] E. R. Hruschka, R. J. G. B. Campello, A. A. Freitas, and A. C. de Carvalho, "A Survey of Evolutionary Algorithms for Clustering," *IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev.*, vol. 39, no. 2, pp. 133–155, 2009.

[104] K. Dutta *et al.*, "Deep learning segmentation of triple-negative breast cancer (TNBC) patient derived tumor xenograft (PDX) and sensitivity of radiomic pipeline to tumor probability boundary," *Cancers (Basel).*, vol. 13, no. 15, p. 3795, 2021.