International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.1 (2025) pp. 1223-1238
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Enhanced hybrid classification model algorithm for medical dataset analysis

N. Kumar[1,2*], T. Christopher[3]

[1]Research Scholar, PG, Research Dept. of Computer Science, Government Arts College, Udumalpet, Tamil Nadu, India
[2]Assistant Professor, Dept. of Computer Science, Dr.N.G.P Arts and Science College, Coimbatore, Tamil Nadu, India
* **Corresponding Author Email:** pkn.kumaar@gmail.com - **ORCID:** 0000-0002-5247-7850

[3]Associate Professor, PG and Research Dept. of Computer Science, Gov.Arts College, Coimbatore, Tamil Nadu, India
Email: chris.hodcs@gmail.com - **ORCID:** 0000-0002-5247-7850

Article Info:

DOI: 10.22399/ijcesen.611
Received : 05 September 2024
Accepted : 25 November 2024

Keywords :

Medical Data Classification (MDC),
Genetic Algorithm (GA),
Convolutional Neural Networks,
Autoencoders (AE),
Outlier Detection (OD).

Abstract:

AbstracThe medical industry generates a significant volume of data that requires effective machine learning models to make accurate predictions for public healthcare. Current Machine Learning (ML) techniques have limitations in feature extraction and classifier accuracy. In this paper using diabetes dataset classification, to address these issues, propose a novel algorithm that enhances Hybrid Classification Model approach by integrating advanced methods tailored for high-dimensional medical data. To handle Missing Values (MV) and outliers, a hybrid imputation approach that combines K-Nearest Neighbor (KNN) and Multivariate Imputation by Chained Equations (MICE) is initially used to preprocess the datasets. Feature extraction (FE) is performed using Deep Feature Extraction techniques, including Convolutional Neural Networks (CNNs) and Autoencoders, followed by Feature Fusion to create a comprehensive feature set. For Feature Selection (FS), introduce an Advanced Ensemble Feature Selection method employing Genetic Algorithm-Based Feature Selection (GAFS), Multi-Objective Evolutionary Algorithm (MOEA), and Relief-Based Methods to identify the most relevant features. Finally, classification is achieved through a Hybrid Classification Model incorporating Ensemble of Classifier with Stacked Generalization (Stacking), Boosting, Bagging and Neural Network (NN) Enhancements with attention mechanisms (AM) and Transfer Learning (TL). This integrated approach enhances the robustness and accuracy of medical data classification. Comparing the suggested approach with current methods, the experimental outcomes show a considerable improvement in accuracy (A), sensitivity (S), specificity (SP), and reduced execution time (ET).

1. Introduction

Accurate diagnosis of medical conditions is an inherently complex task, necessitating a deep understanding and considerable expertise in analyzing diverse medical data. Traditionally, disease diagnosis relies on the examination of medical reports such as Electrocardiograms (ECG), Magnetic Resonance Imaging (MRI), and various diagnostic tests conducted by healthcare professionals. These traditional methods, while valuable, often depend heavily on the expertise of the practitioner and can be limited by subjective interpretation [1]. With the rapid expansion of medical data, there is a growing opportunity to utilize advanced computational techniques to extract meaningful insights and improve diagnostic accuracy. In the modern medical landscape, the availability of extensive clinical datasets and the advancement of data mining methodologies have transformed how we approach disease diagnosis. The use of sophisticated data analysis techniques can uncover hidden patterns within these datasets, significantly aiding clinicians in making more accurate diagnoses [2]. As computerized database systems evolve, they facilitate enhanced decision-making processes and provide the foundation for developing robust diagnostic systems that support clinical decision-making [3].

FE and FS are crucial phases in the medical classification process because they reduce the dimensionality of the data and improve classification accuracy. To control data complexity and preserve the most pertinent information, conventional techniques like Principal (CA) Component Analysis (PCA) and Independent CA

(ICA) have been used [4]. Additionally, the application of feature selection techniques, including evolutionary and heuristic approaches, has proven effective in optimizing data for disease prediction. Techniques inspired by swarm intelligence, known for their efficacy in handling high-dimensional data, are especially beneficial in this context [5].

Despite these advancements, existing algorithms like the Hybrid Random Forest with Back Propagation (HRFBP) still face challenges related to classification accuracy and error rates. For resolving these issues, the research presents a unique algorithm that utilize advanced techniques for pre-processing, FE, FS, and classification. The suggested technique aims to improve the performance of medical data classification (MDC) systems, addressing the shortcomings of current methods and offering improved accuracy and efficiency.

A comprehensive algorithm was established in this study. It surpasses the limitations of existing methods by incorporating cutting-edge techniques in each stage of the data processing pipeline and this technique is considered as main objective of this study. By enhancing pre-processing methods, utilizing sophisticated feature extraction and FS strategies, and refining classification models, the proposed approach seeks to achieve higher diagnostic accuracy and better support clinical (DM) Decision-Making.

The study is structured as follows: The relevant literature on FE, FS, pre-processing, and classification techniques in medical diagnostics is reviewed in Section 2. The new algorithm's suggested methodology is described in depth in Section 3. In Section 4, the performance analysis is discussed and the results of the experiment are presented. A summary of the results and their implications for further research are covered in Section 5.

2. Related Works

MainThe aforementioned problems were suggested to be addressed by Christopher and Kumar (2023) [6]. After the datasets are collected, they are pre-processed using the K-Means Clustering (KMC) technique. Effective management of error rates and missing data is achieved through its utilization. Then, utilizing an Objective Function (OF) to determine the optimum Fitness Values (FV), the features are chosen via the AFOA technique. Selecting the best features using the FV that are most optimal has become the main focus. The medical dataset is then classified using the TBSVM approach. To optimize the margin by a regularization term is the aim of structural risk minimization, or TBSVM. It also reduces training time and improves classification accuracy. It has been revealed that the suggested AFOA-TBSVM procedure works better when compared to the standard methods in terms of higher A, S, and SP as well as faster ET. The experiments' outcomes served as the foundation for these results.

Keywords: Adaptive Firefly Optimization Algorithm (AFOA), FS, Twin Bounded Support Vector Machine (TBSVM) algorithm, and MDC.

1. Introduction: Due to its efficiency and accuracy in MDC, Machine Learning (ML) techniques are being employed more and more for disease detection and diagnosis. Numerous ML algorithms are being utilized to identify different kinds of disorders based on a variety of medical test results. Several diseases could be recognized more rapidly and precisely with the application of these ML techniques [1,2]. Researchers are focusing on developing better ML models for classification in order to more effectively classify data into discrete groups. There are many different input features in these models. Numerous studies are being conducted in this field, yielding novel insights that are profoundly impacting humankind.

Christopher and Kumar (2023) introduced the work uses EFS (Ensemble Feature Selection) with BPNN (Back Propagation Neural Networks) to handle the afore mentioned issues [7]. The input data is pre-processed using KMC (K-Means Clustering) algorithm, mainly for handling missing values and subsequently, EFS method is used to choose the features since it produces the best FV using an OF. To solve the FS problem, EFS relies on integrating many FS rather than just one FS. Combining the results of multiple single FS approaches, such as EEHO (Entropy Elephant Herding Optimisation) and AFOA (Adaptive Firefly Optimisation Algorithm), is one alternative for the EFS method. And EBFO (Entropy Butterfly Optimization Algorithm) acquire improved outcomes rather than utilizing a single FS methodology. Finally, the medical dataset classification is performed using BPNN algorithm. With the help of the BPNN algorithm, a multilayer FFNN (feed forward neural networks) is trained. The class labels in tuples are predicted using weights that are learnt iteratively. The experimental findings of the proposed EFS-BPNN algorithm demonstrates better values for accuracy, sensitivity, specificity, and execution time when compared with existing methods.

Christopher and Kumar (2023) introduced Hybrid Random Forest with Back Propagation (HRFBP) neural network algorithm is proposed [8]. Initially, the datasets are collected which is preprocessed using K-Means Clustering (KMC) algorithm. Error

rates and MV are well managed by it. Then, the feature extraction (FE) is done by using Modified PCA (MPCA) which is focused to extract the significant features from the given medical dataset. After that, the FS is done by using EFS algorithm which generates best fitness values via objective function. EFS is done dependent on integrating numerous FS rather than a single FS to handle the FS issue. The possibility of EFS method is that merging the outcomes of a various single FS methods like Entropy Elephant Herding Optimization (EEHO), Adaptive Firefly Optimization Algorithm (AFOA) and Entropy Butterfly Optimization Algorithm (EBFO) acquire improved outcomes rather than utilizing a single FS methodology. Finally, the medical dataset classification is performed using HRFBP algorithm. The HRFBP algorithm performs training and testing process which learns a set of weights for prediction for the class label of features. HRFBP increases the classifier accuracy and reduces the error rates prominently. Based on the testing results, it was determined that compared to the current algorithms, the suggested HRFBP algorithm performs better in terms of higher A, S, SP, and shorter ET.

Modern techniques and models for missing data imputation were put to the test and improved in [9] by Psychogyios et al. (2023). For the pre-imputation problem, we suggest a novel MV Imputation (MVI) technique based on demonising autoencoders (DAE) and KNN. In order to get more accurate results, optimize the training process by reapplying KNN to the missing data every N epochs, each time using a different value for the variable k. A GAN-based State-Of-The-Art (SOTA) method for imputation of missing data is also revised. Then, introducing enhancements to the architecture and training process using this as a baseline. These models are compared to the ones that are typically used for both post-imputation prediction and imputation tasks in clinical research investigations. The outcomes demonstrate that the suggested Deep Learning (DL) techniques perform better than the current baselines, producing more accurate predictions and better imputation.

Shang et al. (2021) in [10] have proposed an Information Entropy (IE) based multi-scale deep feature fusion (MSFF) and Intelligent Fault Detection (IFD) technique. To create a multi-scale Deep NN (DNN) FE structure, a standard AE, demonising AE, sparse AE, and contractive AE are first applied in parallel. To achieve (LD) low-dimensional features, the stability of the framework was ensured, and preserve the correctness of the deep features, an IE-based DFF technique is recommended. Deep Belief Network (DBN), a potential framework and a fault classifier for FD. Utilizing a gearbox test-bed, the suggested method's efficacy was confirmed. The experimental results show that compared to current and conventional intelligent FD approaches, the recommended method can more accurately classify the raw data by extracting pertinent information and features.

Units comprising data conversion, data pre-processing, normalization, FE, dataset splitting, and classification and prediction unit are presented by Shaikh et al (2024) [11]. The suggested prediction approach's unique characteristics include its capacity to categorize the type of disease based on a patient's medical report , and also identify vector-borne disease in initial stage. While seven distinct forms of traditional machine learning (ML) and one Hybrid ML (HML) are employed for classification, Recurrent Neural Network (RNN) based Reinforcement Learning (RL) is utilized for recommendation. A series of experiments were conducted to assess the efficiency of the suggested approach.

A set of 1539 distinct cases of a vector-transmitted disease has been compiled into a dataset. For experimental assessment, eleven prevalent vector-borne diseases, including chikungunya, dengue, Japanese encephalitis, kala-azar, and malaria, were selected. The suggested prediction model's performance accuracy is 98.76%, which helps the medical team make decisions quickly and finally contributes to the patient's survival. The algorithm in the final step offers recommendations for the classifiers that provide four distinct classes, namely normal, mild, moderate, and severe. The suggestion also outlines the path forward for treating vector-borne diseases in the future.

The prediction of abdominal disorders, such as renal and liver diseases, was suggested by Vijayarani et al. (2023) in [12]. Using classification algorithms, the work seeks to predict liver disorders such as acute hepatitis, cirrhosis, bile duct, chronic hepatitis, and liver cancer. Using classification algorithms, acute glomerulonephritis, acute nephritic syndrome, acute renal failure, and chronic kidney disease are among the kidney disease that the research aims to predict. In order to predict liver and renal disorders, this work suggests a novel hybrid classification technique called WRFSVM (Weighted Random Forest SVM).

Using two datasets for diabetes and heart attacks, as well as two types of cancer (lung and breast), Lafta et al. (2019) presented an approach for identifying a number of diseases in [13]. The role of classifier was played by BPNN (Back Propagation (NN) Neural Network). NN performance is enhanced by using the (GA) Genetic Algorithm, which provides the classifier with the optimum features to

maximize the classification rate. Based on the quantity, features, and kind of data, the system demonstrated significant efficiency in handling databases that differ from one another. The results supported this, with the majority of the datasets showing a classification ratio of 100%.

A novel ML strategy to predict cardiac disease has been suggested by Kavitha et al. (2021) in [14]. Regression and classification are two Data Mining (DM) approaches that are applied in the suggested research, which makes use of the Cleveland heart disease dataset.Then Decision trees (DT) and RF are the ML techniques thar are used. The ML model's advanced method is established. Three ML techniques are employed in the execution: 1. RF, 2. DT, and 3. Hybrid model (a hybrid of DT and RF). Based on testing results, an accuracy level of 88.7% in predicting heart disease (HD) was attained by the hybrid model. In order to predict HD, an interface design hybrid model consisting of DT and RF was employed to gather user input parameters.

3. Proposed Methodology

Hybrid Classification Model method is suggested in this research to improve performances of medical dataclassifications using four processes namely pre-processing, feature extraction, feature selections, and classifications. Figure 1 displays the proposed Hybrid Classification Model overall flow diagram.

3.1 Input Data Collection

This work's datasets obtained from UCI ML repository included the datasets: Pima Indians with diabetes; Heart diseases and hepatitis, and fertility. The Pima databases contain outputs of patients with pregnancies, BMI, insulin levels, ages, glucose levels, blood pressures, skin thicknesses, and family histories of diabetes, together with counts of medical predictors (independent) factors.

3.2 Hybrid Imputation Method for preprocessing

This paper provides efficient methods for imputing different patterns of missing data by presenting a six-layer hybrid model that hybridizes few imputation strategies [15]. As presented in Figure 2 is the framework that is being given consists of Analysing the Original Dataset -The initial phase involves a thorough examination of the dataset to identify different missing data patterns. This analysis helps in understanding the nature of the missing data, whether it's completely random, non-random, or random. Decomposing the Dataset - After identifying the missing data patterns, the next step is to decompose the dataset based on these patterns. This involves separating the dataset into distinct subsets, each representing a specific type of missing data pattern. Imputing Data in Each Subset - Each subset from the decomposition step is then imputed using the most suitable techniques. This may involve a combination of different imputation methods, such as MICE or KNN, to estimate the MV. The features of the data in each subset determine the strategy to be used. Merging Imputed Data - Once the best possible estimations for the MV are obtained for each subset, these are combined to form a complete, imputed dataset. This final dataset is expected to have a higher data quality and be more suitable for further analysis or modelling. This multi-layer method attempts in improving the accuracy and reliability of data imputation, especially when dealing with datasets containing complex missing data patterns. It ensures that different types of missing data are addressed using the most appropriate methods, improving overall data integrity and usefulness for subsequent analyses,

1. Examining missing data patterns: This stage involves analyzing the initial dataset that contained a large amount of missing data in order to find the (MP) missing patterns.
2. Imputing missing data through MNAR pattern: The MNAR pattern-exhibiting attributes were located, and the relevant constant global label values were used to impute their missing values.
3. Decomposing: By determining the internal relationships between the variables, the reason for this missing, and relevant descriptions of the MP, as well as conferring with an endocrinologist. Two datasets with MCAR and MAR patterns, DMCAR and DMAR, are created from the imputed dataset from Step 2.
4. Single imputing: KNN is one of the single imputation approaches [16].
5. In the DMCAR dataset, the MCAR patterns are imputed using (KNN) and hot-desk. The best results are then chosen as the WinnerD_{MCAR}, and the outcomes from each imputation technique are evaluated using several classifiers.
6. Multiple imputing (MI): The DMAR datasets with the MAR pattern were subjected to numerous imputation techniques, such as expectation-maximization (Em), multivariate imputation by chained equations (MICE), and Markov chain Monte Carlo (MCMC). This study produced five distinct datasets with every multiple imputation technique. Researchers compared the results of multiple classifiers to assess the datasets imputed by every MI technique. The WinnerD_{MAR} is then selected based on the best outcomes.

7. Hybrid imputation: The last stage creates the final dataset by combining the WinnerDMCAR and WinnerDMAR datasets that were chosen in steps 4 and 5. Repetitive features are then removed.

- **Missing Completely at Random (MCAR) Technique:** Using the MCAR technique, creating an algorithm that can randomly remove values in order to normalize the dataset is the most practical way to handle missing values.
- **Missing at Random (MAR) :**The values of missing variables are found in a subset of the X dataset, not in the entire dataset Y. This indicates that the MAR approach employs covariance to find the missing values by carefully examining the larger dataset to find the values that don't match the variable being studied.
- **Missing Not at Random (MNAR) :**The frequency of MV becomes reliant on b or other unanticipated factors in the dataset due to a violation of the requirements for managing the missing values based on the MAR approach. For example, MNAR can be used to infer from tax computations that the values of missing data are contingent upon tax payers' unobserved revenue disclosures. The benefit of the MNAR technique is its ability to distinguish between data that was never provided and data that was entered wrongly because of measurement error.

The proposed six-layer hybrid imputation model effectively addresses complex missing data patterns through a systematic and multi-faceted approach. By analyzing, decomposing, and imputing data according to identified patterns namely MCAR, MAR, and MNAR, the model ensures that the most appropriate imputation techniques are applied. Accurate estimation and reconstruction of missing values are made easier by the application of multiple imputation methods like MCMC and MICE as well as single imputation approaches like KNN. The final hybrid imputation step integrates the results to produce a comprehensive, high-quality dataset, enhancing data integrity and suitability for further analysis [17]. This approach not only improves the reliability of imputed datasets but also provides a robust framework for managing missing data and outliers effectively.

3.3 Deep Feature Extraction using CNN's and Autoencoders

Deep FE using CNN and Autoencoders involves leveraging CNNs to learn spatial hierarchies of features from images and Autoencoders to compress data into a lower-dimensional representation. This combination is effective for capturing essential data characteristics, making it valuable for applications like image recognition, noise reduction, and anomaly detection. CNNs focus on extracting meaningful patterns, while Autoencoders compress and reconstruct these features, enabling efficient and accurate data analysis.

CNN's (Convolutional Neural Network)

Applications for image-based learning frequently make use of CNNs. CNNs can extract useful information from training data by using an automatic FE approach [18]. In CNN framework, a large number of convolutional Layer (CL), pooling Layer, and fully connected (FC) layers are commonly utilized. Convolutional kernels are used to convolve the input for FE , as seen in Figure 3. The pooling layer preserves the resolution of the Feature Map (FM) while reducing the computational volume of the network. In CNNs, as the number of layers increases, the pooling layer size typically decreases. Two of the most used types of pooling layers are maximum and average pooling.

AutoEncoders (AEs)

AEs are classified as unsupervised learning algorithms. Since an AE does not need labeled data for training, in short, it compresses input data to a LD latent space before reconstructing it by decompressing the latent space representation. When it comes to reducing dimensionality during the compression step, AEs are similar to PCA [19].AEs, unlike PCA, do nonlinear transformations utilizing deep neural networks. Figure 4 illustrates the construction of a typical AE. In order to create a latent (hidden) space with fewer dimensions than the original input, high-dimensional (HD) input data is encoded (compressed). To obtain decompressed findings, the latent representation is reconstructed, or decoded. In finally, deep feature extraction using Convolutional Neural Networks (CNNs) and Autoencoders combines the strengths of both approaches to enhance data analysis. CNNs excel at automatically learning spatial hierarchies and extracting meaningful patterns from image data, making them ideal for applications such as image recognition and object detection. On the other hand, AE are useful for tasks like noise reduction and anomaly detection because of their capacity to do nonlinear (DR) Dimensionality Reduction. This means that AE may compress HD input into a LD latent space and reconstruct it. By leveraging CNNs for detailed feature extraction and Autoencoders for efficient data compression and reconstruction, this

combined approach offers a powerful framework for accurate and scalable data analysis across various domains.

3.4 Advanced Ensemble Feature Selection using GAFS, MOEA and Relief-Based Methods

The Advanced Ensemble Feature Selection method combines three powerful techniques to identify the most relevant features in a dataset. It employs Genetic Algorithm-Based Feature Selection (GAFS), which uses evolutionary principles to simulate natural selection for feature selection, MOEA to optimize multiple criteria simultaneously for a balanced feature set, and Relief-Based Methods to assess the significance of features based on their ability to distinguish between similar instances. By integrating these methods, the approach aims to improve the effectiveness of feature selection compared to using any single method alone. 3D Brain Tumor Detection with Hybrid Mean Clustering and Ensemble Classifiers was optimised [20].

Genetic Algorithm Based Feature Selection (GAFS)

Among the most sophisticated FS algorithms is the GA. Natural genetics and the physics of biological evolution serve as the foundation for this stochastic function optimization technique. In order to better adapt to their environment, organisms' genes typically develop throughout generations. Figure 5 [21] presents a state chart diagram of a GA-based FS process. Operators like initialization, fitness assignment, crossover, mutation, and selection make up a GA. The operators and parameters of the GA are then discussed in detail.

Initialization operator: Establishing and initializing each member of the population is the first stage. As a stochastic optimization technique, GA randomly initialize an individual's genes.

Fitness assignment operator: After initialization, there is a need to give each member of the population a fitness value. Testing data is used to evaluate each neural network's performance after it has been trained using training data. Poor fitness is indicated by a large selection error. Recombination is more likely to choose individuals with greater fitness. A rank-based fitness assignment method was used in this study to determine the FV of each participant. Selection operator (SO): Once an individual has completed a fitness task, a SO is used to select them to work in the recombination for the next generation. High fitness individuals are able to live in the environment. Researchers chose people using the stochastic sampling replacement technique, where the weight of the elements determines an individual's fitness. With N being the population size, N/2 is the number of individuals chosen. Crossover Operator (CO): After half of the population has been chosen by the SO, CO operators are utilized to create a new population. This operator creates children for the new population by randomly choosing two individuals and combining their traits. Whether a child's traits are inherited from one or both parents are determined via the uniform CO approach. Mutation Operators: Remarkably similar offspring can be produced by the CO operator. The mutation operator, which modifies some of the offspring's traits at random, provides a solution to this issue. We create a random number between 0 and 1 to determine whether a feature has been altered.

Multi – Object Evolutionary Algorithm (MOEA)

Because it fail to consider many solutions, traditional mathematical programming techniques like calculus become challenging when attempting to solve multi-objective optimization issues. Rather, under these circumstances, population-based Meta-Heuristic (MH) methods like Evolutionary Algorithms (EA) work better. Many academics have developed different implementations of the MOEA after Goldberg proposed the notion in 1989 utilizing the idea of domination. MOEAs have changed over time, starting with conventional aggregating techniques and moving on to elitist models of Pareto-based algorithms in the late 1990s and indicator-based algorithms more recently. Notwithstanding its absence of elitism, popular Pareto-based techniques such as multiple objective GA, niched Pareto GA (NPGA), and non-dominated sorting GA (NSGA) [22] were evaluated on several real-time applications and did not always preserve non-dominated solutions. Then, elitist MOEAs were created to deal with this problem, including SPEA, SPEA2, PAES, PESA, PESA-II, and NSGA-II. These Pareto-based elitist methods are frequently applied in the most recent MOEA applications for FS problems. Using a selection mechanism based on performance metrics, such as the indicator-based EA, is a contemporary trend in MOEA design. The Elitist Pareto-based MOEA for Diversity Reinforcement (ENORA - $\mu+\lambda$) method, which ranks population members according to their non-domination degree, has recently been merged with the NSGA-II.

Relief –Based Methods

One of the most widely used filter-based FS techniques is a member of an algorithmic family that all use the same fundamental ideas in feature selection. This family of algorithms, often referred to as Relief-based algorithms, is founded on the notion that a feature's local relevance should be

assessed in relation to the context that the other data provide [23]. This is achieved by measuring each feature's relevancy univariate, while selecting samples for the feature's univariate analysis based on neighborhoods defined by all the characteristics in the observed data and increasing the robustness of feature quality estimation with noisy data (Figure 6). The Advanced Ensemble Feature Selection method, which integrates Genetic Algorithm-Based Feature Selection (GAFS), Multi-Objective Evolutionary Algorithm (MOEA), and Relief-Based Methods, offers a comprehensive and effective approach to feature selection. GAFS utilizes evolutionary principles to iteratively refine feature subsets, MOEA optimizes multiple criteria to ensure a balanced feature set, and Relief-Based Methods evaluate the significance of features based on their ability to differentiate between instances. The ensemble approach offers a more efficient solution than any one method alone by combining different strategies to improve the precision and resilience of FS. In the end, this integrated technique produces more accurate and perceptive Data Analysis (DA) by enhancing the quality and relevance of selected features.

3.5. Hybrid classification Model

Classification is achieved through a Hybrid Classification Model incorporating Ensemble of classifier with Stacked Generalization (Stacking), Boosting, Bagging and Neural Network enhancement with attention mechanisms, and Transfer Learning. This integrated approach enhances the robustness and accuracy of medical data classification.

Ensemble Classifier

Using several classifiers or models in combination to improve classification efficiency over a single classifier is known as an ensemble classifier. Displayed in Figure 7 Several ensemble-based classification approaches include stack-based ensemble, boosting, RF, and bagging [24]. Building numerous models of the same type on various subsamples of the same training dataset and combining their results is known as bagging. Boosting also involves creating a chain of identical models, each of which learns to decrease the prediction error of the preceding model. Several Meta classifiers are used in the current study to implement stack-based ensemble classification. Using various Meta learners, many models of various types are stacked in a process known as stacking classifiers, which is ensemble-based classification.Meta learners are trained on the output of these base classifiers to aggregate their results as best they can once the base classifiers have been trained on all of the training data. In order to aggregate the strength of all the base models, heterogeneous classifiers are employed to form an ensemble at the base level. The efficiency of a stack-based ensemble outperforms the best base layer classifier in the majority of cases.

Bagging

Bootstrap aggregating, another name for the bagging technique, is a procedure that is entirely dependent on the data. It describes the process of extracting several tiny subsets of data from the original dataset. By altering the stochastic distribution of the training dataset, which varies the model's predictions greatly with small changes to the training set. The goal of bagging is to generate more varied prediction models. Aggregating and bootstrapping together is known as "bagging". The training dataset in bootstrapping is replicated during the ensemble model training process. The ultimate outcome in aggregation is determined by a majority vote of the model's predictions [25], which are then used to generate the final forecast. The benefit of bagging is that it lowers variance, which gets rid of overfitting. It functions nicely with high-dimensional data as well.

Boosting

Every new model in this sequential process seeks to address errors in the previous one. Boosting is the process of fitting progressively more weak learners in a highly flexible way. Every model in the series is fitted, and weight is added to dataset observations that the frameworks in the previous sequences handled poorly. Similar to bagging, boosting can be applied to problems involving regression and classification. The three types of boost algorithms are: Stochastic GB (SGB), Adaptive Boosting (AdaBoost), and Extreme GB (XGB), also called XGBoost.Numerous research has used different kinds of boosting. The AdaBoost algorithm, for instance, is used in voice feature extraction and noise detection. The classification of fake news uses the XGB algorithm.

Stacking

Data from many predictive models are combined to create a novel framework (meta-model) using the stacking strategy, also known as stacked generalization. A meta-model, also called a level-1 model, which integrates the predictions of the base models, and two or more base models, sometimes called level-0 models, make up the architecture of a stacking model. Bone Cancer Diagnosis through Deep Learning on Medical Imagery was studied [26]. Level zero, or base, models are those that fit

on training data and generate predictions. The level 1 model, or meta-model, is where the model learns the way to combine the predictions of the fundamental models in the most efficient way. Class labels or probability values are examples of the outputs from the base models that are input into the meta-model in the classification event. Usually, the stacking technique works.

Neural Network Enhancements

In order to improve performance in tasks like machine translation, AM allow NN to dynamically focus on the most relevant portions of the input data. Transfer learning (TL) boosts neural network capabilities by leveraging pre-trained models on similar tasks, significantly reducing the need for extensive data and training time. Finally, advanced voting mechanisms, such as ensemble methods, combine the predictions of multiple models to increase overall accuracy and robustness, leading to better generalization and performance on diverse datasets.

Attention Mechanisms: When a model is making predictions or producing output, it can choose focus on specific portions of its input due to a DL technique called the AM. In order to enhance their performance on tasks like Object Detection (OD) and image restoration,AM are also employed in Computer Vision (CV) models. Channel attention and self-attention are the two common attention processes at the moment [27]. Depending on various channels of attention, the channel AM in the NN selectively emphasizes or suppresses specific channels of the FM. For instance, to significantly increase network performance, incorporate the channel AM into the residual block of the super-resolution network RCAN. A residual non-local attention network was presented as a mechanism for high-quality image restoration. The purpose of the self-AM, which is frequently employed in models like Transformer and non-local NN, is to apply self-attention to FM in order to selectively focus on distinct spatial areas in the input image.

Advanced Voting Mechanism: Enhances decision-making in ensemble classifiers by considering not only the majority vote but also the confidence scores and prediction probabilities. Each base classifier provides a confidence score or probability for its prediction [28]. A more sophisticated voting system aggregates these scores to make final decisions. Leads to more accurate and reliable predictions by taking into account the certainty of each classifier's prediction.

TL: Utilizes pre-trained models on similar tasks and fine-tunes them on the target dataset. Pre-trained models (e.g., on large datasets like ImageNet) are adapted to specific tasks (e.g., medical imaging). Requires less data and training time [29]. Leverages existing knowledge to improve performance on the target task with minimal additional training. In finally Figure 8 is the Hybrid Classification Model Combines Ensemble Methods, Neural Network Enhancements, and Transfer Learning to create a highly effective classification system for medical data. Ensemble methods like Bagging, Boosting, and Stacking improve accuracy and robustness by leveraging multiple classifiers and combining their predictions. Neural network enhancements, including attention mechanisms and advanced voting systems, refine the model's focus and decision-making. Transfer learning adapts pre-trained models to specific tasks, reducing the need for extensive data and training. This integrated approach enhances classification performance, increases generalization across diverse datasets, and ensures more reliable and accurate results.

4. Experimental Results

To evaluate how well the Hybrid Classification Model performed, four data sets from the UCI ML repository were used [30]. Input data for ML were normalised between [0, 1]. The existing methods used pre-processing for Hybrid Imputation methods combining MICE and KNN's, Feature extraction combining CNN and Autoencoders, Feature selection employing GAFS, MOEA and Relief-Based methods which is evaluated with Hybrid Classification Model algorithm. The suggested and current algorithms are compared using performance metrics like execution time, sensitivity, accuracy, and specificity. The outcomes of the performance comparisons are shown in Table 1. The Pima data is gathered from the reference [31]

Accuracy

By dividing the $(T_p + T_n)$, the total actual classification parameters to $(T_p + T_n + F_p + F_n)$, he sum of the classification parameters, accuracy is calculated. This accuracy is defined as the overall accurateness of the framework.

$$\text{Accuracy} = \frac{T_p+T_n}{(T_p+T_n+F_p+F_n)} \quad (1)$$

Where Tp is true positive,
Tn is true negative,

Fp is false positive,
and Fn is false negative
The Hierarchical Clustering Model (HCM) achieved the best accuracy at 92.17%, suggesting higher classification capabilities, according to the performance analysis of the approaches for classifying diabetes data, as shown in Figure 9. The Hybrid Random Forest with Back Propagation (HRFBP) method followed with an accuracy of 89.33%, demonstrating strong performance. The Enhanced Feature Selection - Back Propagation Neural Network (EFS-BPNN) and Adaptive Feature Optimization Algorithm - Twin Bounded Support Vector Machine (AFOA-TBSVM) methods achieved accuracies of 87.45% and 84.89%, respectively, with AFOA-TBSVM showing the least accuracy. This comparison underscores the effectiveness of hierarchical clustering and hybrid methods in enhancing accuracy in medical dataset classification.

Recall

Recall or sensitivity, often referred to as the TP rate, recall, or probability of detection in different measurement domains, is the proportion of TP that are correctly identified as true. As demonstrated in equation 2.

$$\text{Recall} = \frac{T_p}{T_p + F_n} \quad (2)$$

The performance analysis graph in Figure 10, compares the recall percentages of four classification models for diabetes prediction: AFOA-TBSVM, EFS-BPNN, HRFBP, and HCM. The Hierarchical Clustering Model (HCM) demonstrates the highest recall at 91.56%, indicating superior sensitivity in identifying true positive cases. The Hybrid Random Forest with Back Propagation (HRFBP) method follows with a recall of 88.42%, showing strong performance but still trailing behind HCM. The Enhanced Feature Selection - Back Propagation Neural Network (EFS-BPNN) model achieves a recall of 86.23%, which is respectable but lower than the top two models. The AFOA-TBSVM model has the lowest recall at 85.11%, indicating the least effective performance among the four. This comparison highlights the HCM's robustness in accurately identifying diabetes cases, outperforming the other evaluated models.

Precision

Precision refers to the percentage of the results which are relevant and defined as equation 3,

$$\text{Precision} = \frac{Tp}{tp + fp} \quad (3)$$

Figure 11 is the performance analysis graph compares the precision percentages of four classification models for diabetes prediction: AFOA-TBSVM, EFS-BPNN, HRFBP, and HCM. The Hierarchical Clustering Model (HCM) achieves the highest precision at 89.72%, indicating a superior ability to correctly identify positive cases among those classified as positive. The Hybrid Random Forest with Back Propagation (HRFBP) method follows with a precision of 87.29%, showing strong but slightly lower performance than HCM. The Enhanced Feature Selection - Back Propagation Neural Network (EFS-BPNN) model has a precision of 86.45%, which is respectable but not as high as the top two models. The AFOA-TBSVM model has the lowest precision at 83.2%, indicating the least effective performance in minimizing false positives. This analysis underscores the HCM's effectiveness in ensuring a high precision rate, outperforming the other evaluated models in correctly predicting diabetes cases.

F – measure

The feature rating is obtained; the higher the score, the more essential the feature. The F-measure analysis graph for diabetes prediction is shown in Figure 12. The performance analysis graph compares the F-measure percentages of four classification models for diabetes prediction: EFS-BPNN, HRFBP, AFOA-TBSVM, and Hierarchical Clustering Model HCM. The best balance between recall and precision is demonstrated by the HCM, which obtains the greatest F-measure of 88.12%. The Hybrid Random Forest with Back Propagation (HRFBP) method follows with an F-measure of 86.79%, demonstrating strong but slightly lower performance than HCM. The Enhanced Feature Selection - Back Propagation Neural Network (EFS-BPNN) model has an F-measure of 85.14%, which is respectable but not as high as the top two models. The AFOA-TBSVM model has the lowest F-measure at 84.14%, indicating the least effective performance among the evaluated models. This analysis underscores the HCM's effectiveness in achieving a high F-measure, outperforming the other models in balancing precision and recall for diabetes prediction.

Execution Time

When the suggested algorithm runs in less time, the system performs better. As illustrated in the Figure 13, The performance analysis graph compares the execution times of four classification models for diabetes prediction: EFS-BPNN, HRFBP, AFOA-TBSVM, and HCM. The Hierarchical Clustering Model (HCM) achieves the shortest execution time at 12.83%, indicating superior efficiency and faster processing speed. The Hybrid Random Forest with

Back Propagation (HRFBP) method follows with an execution time of 15.19%, demonstrating strong but slightly slower performance than HCM. The Enhanced Feature Selection - Back Propagation Neural Network (EFS-BPNN) model has an execution time of 17.62%, which is respectable but not as quick as the top two models. The AFOA-TBSVM model has the longest execution time at 19.45%, indicating the least efficient performance among the evaluated models. This analysis underscores the HCM's effectiveness in ensuring a high efficiency rate, outperforming the other models in terms of processing speed for diabetes prediction.

4. Conclusions

In conclusion, this study presented a new procedure to enhance the Hybrid Classification Model for high-dimensional medical data, with a specific application to diabetes dataset classification. The

***Table 1**. Outcomes of the comparisons of the performances*

Metrics and dataset	Methods			
	EFS-BPNN	HRFBP	AFOA-TBSVM	HCM
Accuracy - Diabetes	87.45	89.33	84.89	92.17
Recall - Diabetes	86.23	88.42	85.11	91.56
Precision - Diabetes	86.45	87.29	83.20	89.72
F-measure -- Diabetes	85.14	86.79	84.14	88.12
Execution time - Diabetes	17.62	15.19	19.45	12.83

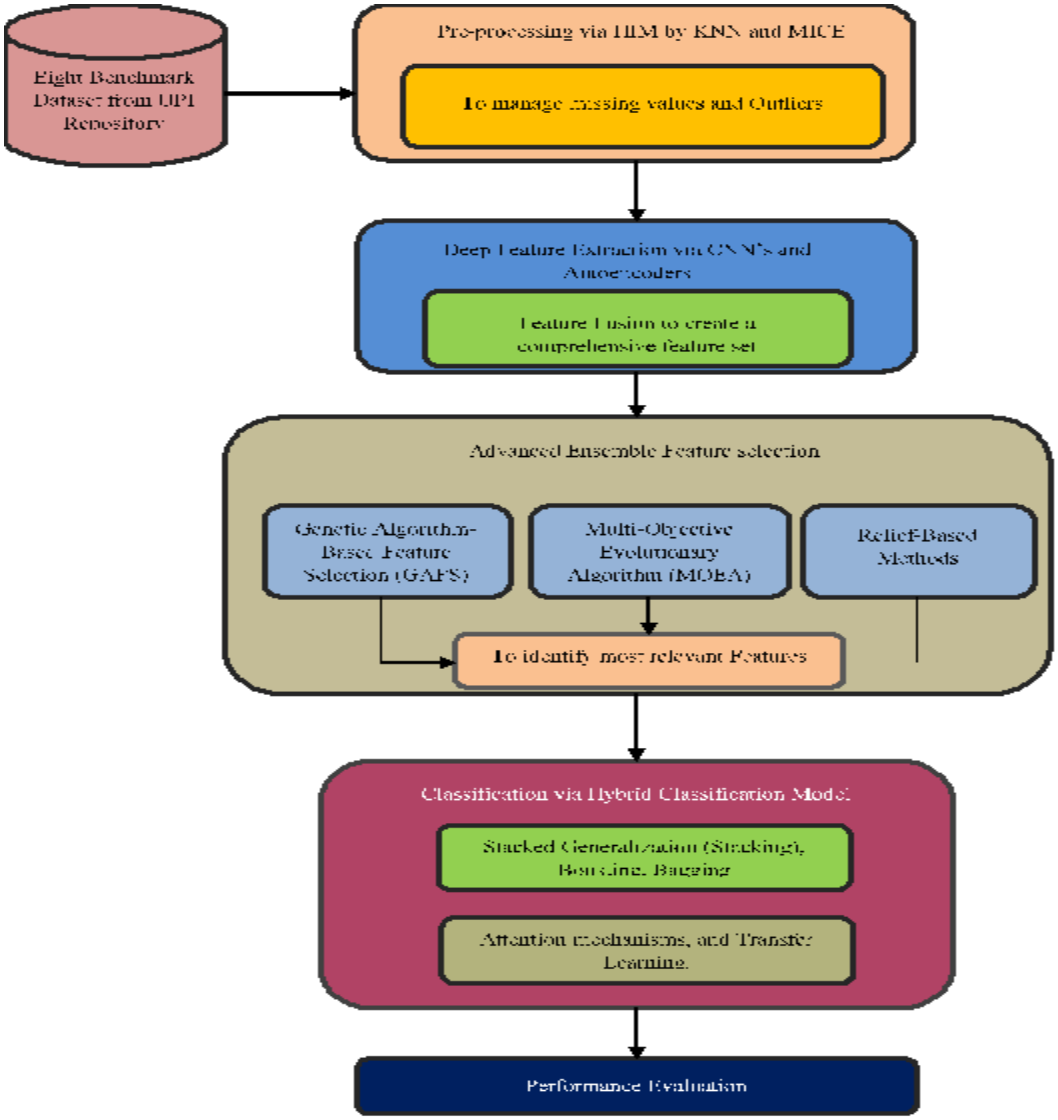

***Figure 1.** Overall block diagram of suggested Hybrid Classification Model*

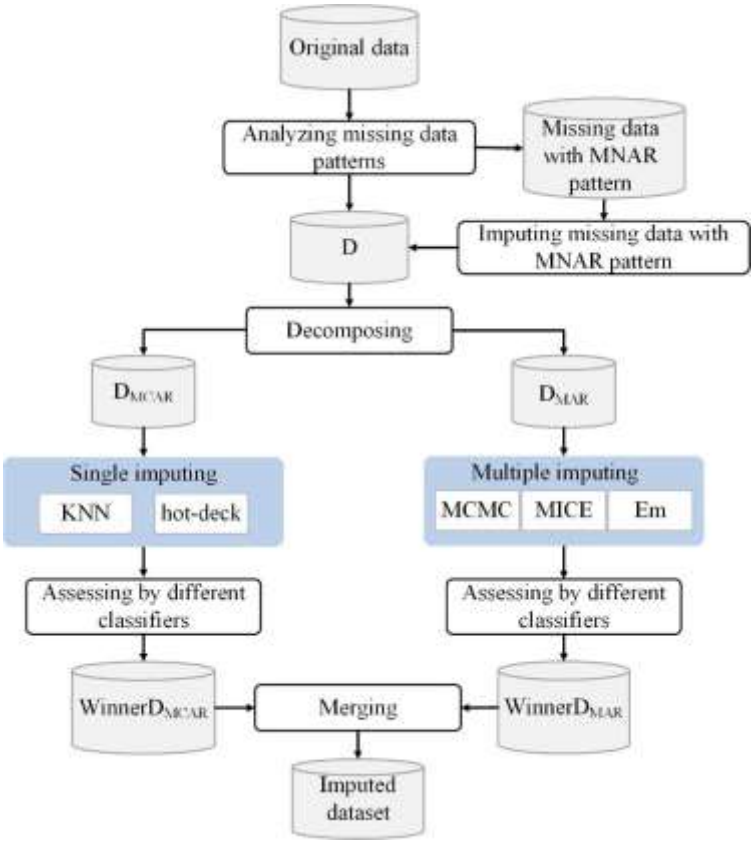

Figure 2. *Hybrid Imputation model for missing Data*

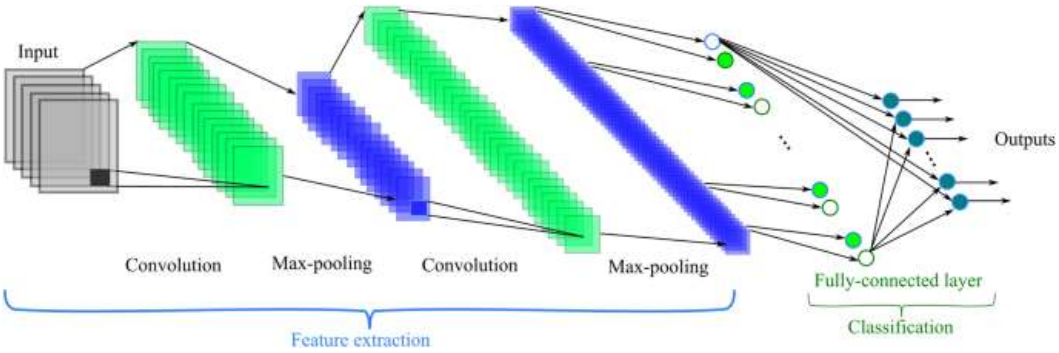

Figure 3. *CNN's Schematic*

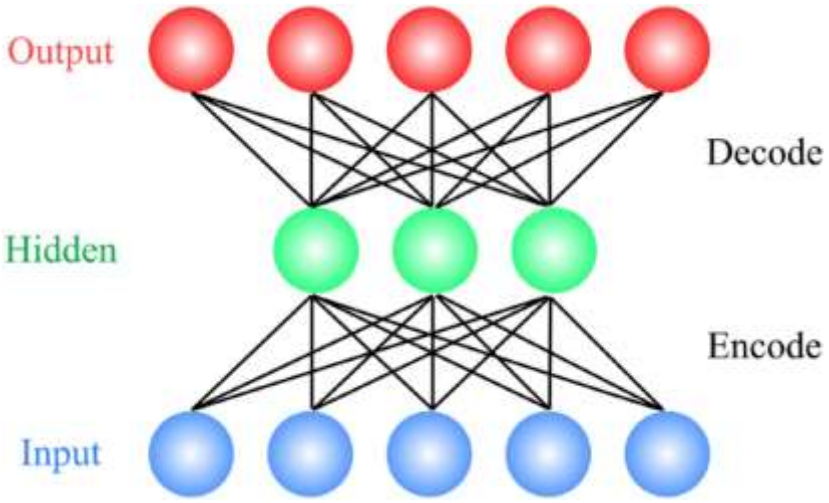

Figure 4. *AE framework*

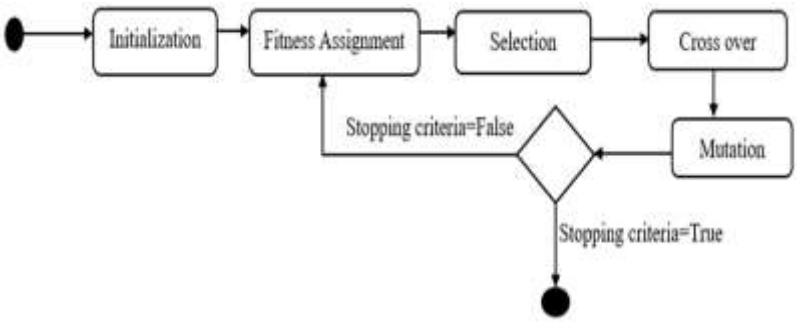

Figure 5. *Genetic algorithm feature selection*

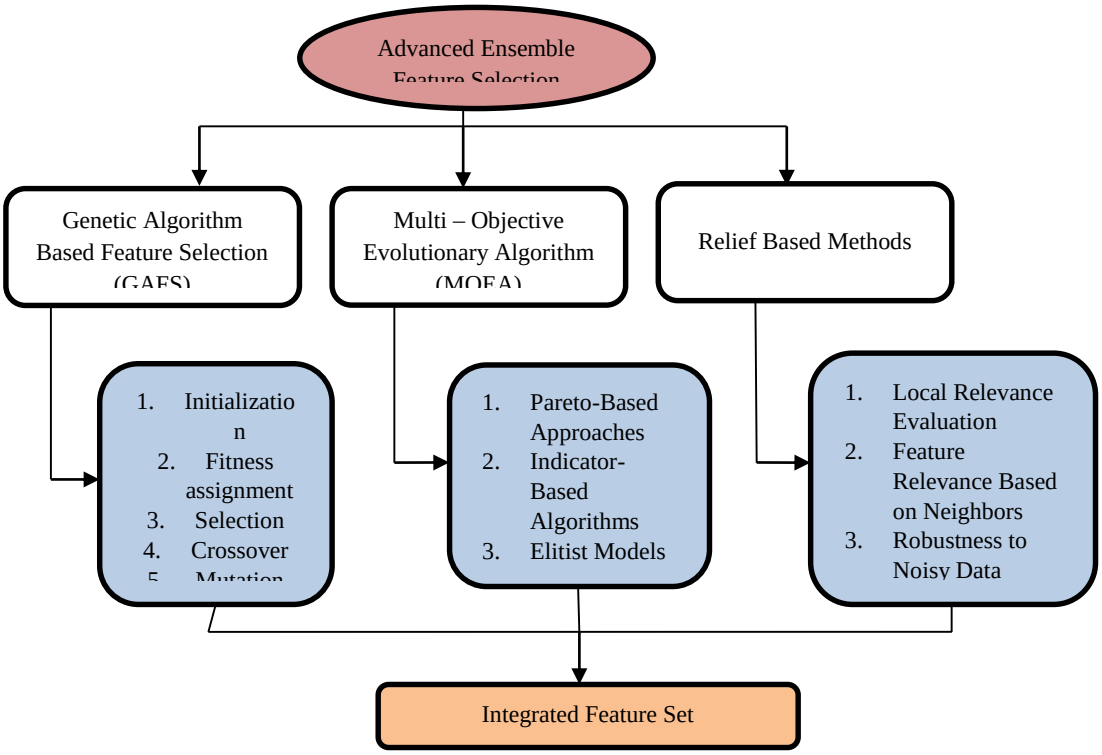

Figure 6. *Advanced Ensemble Feature Selection Flow Diagram*

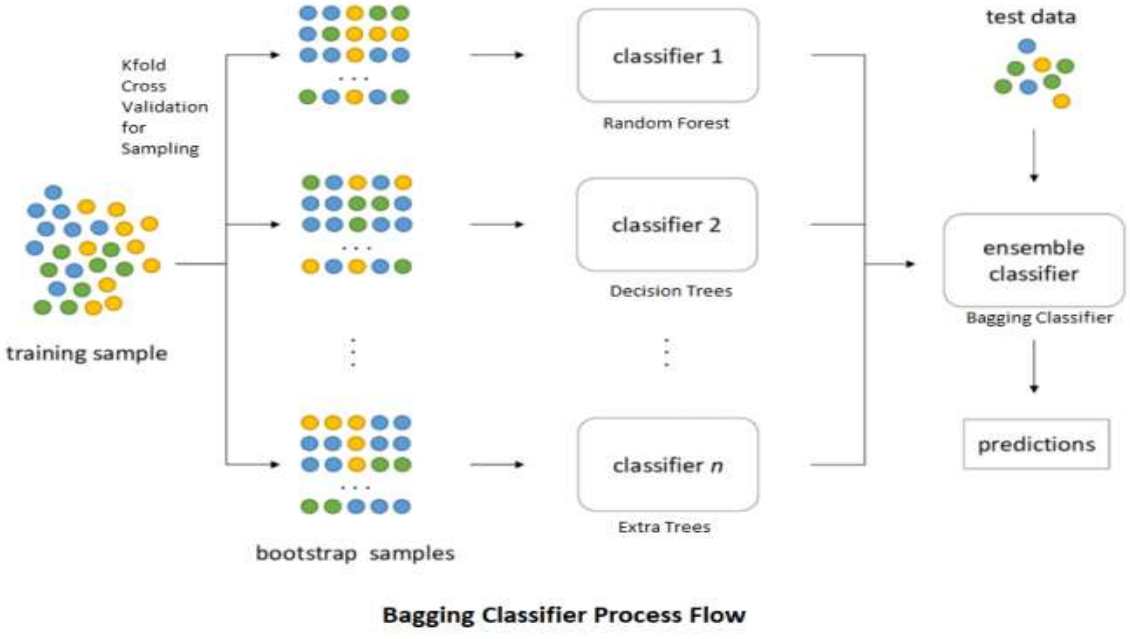

Figure 7. *Ensemble Classifier Basic Process diagram*

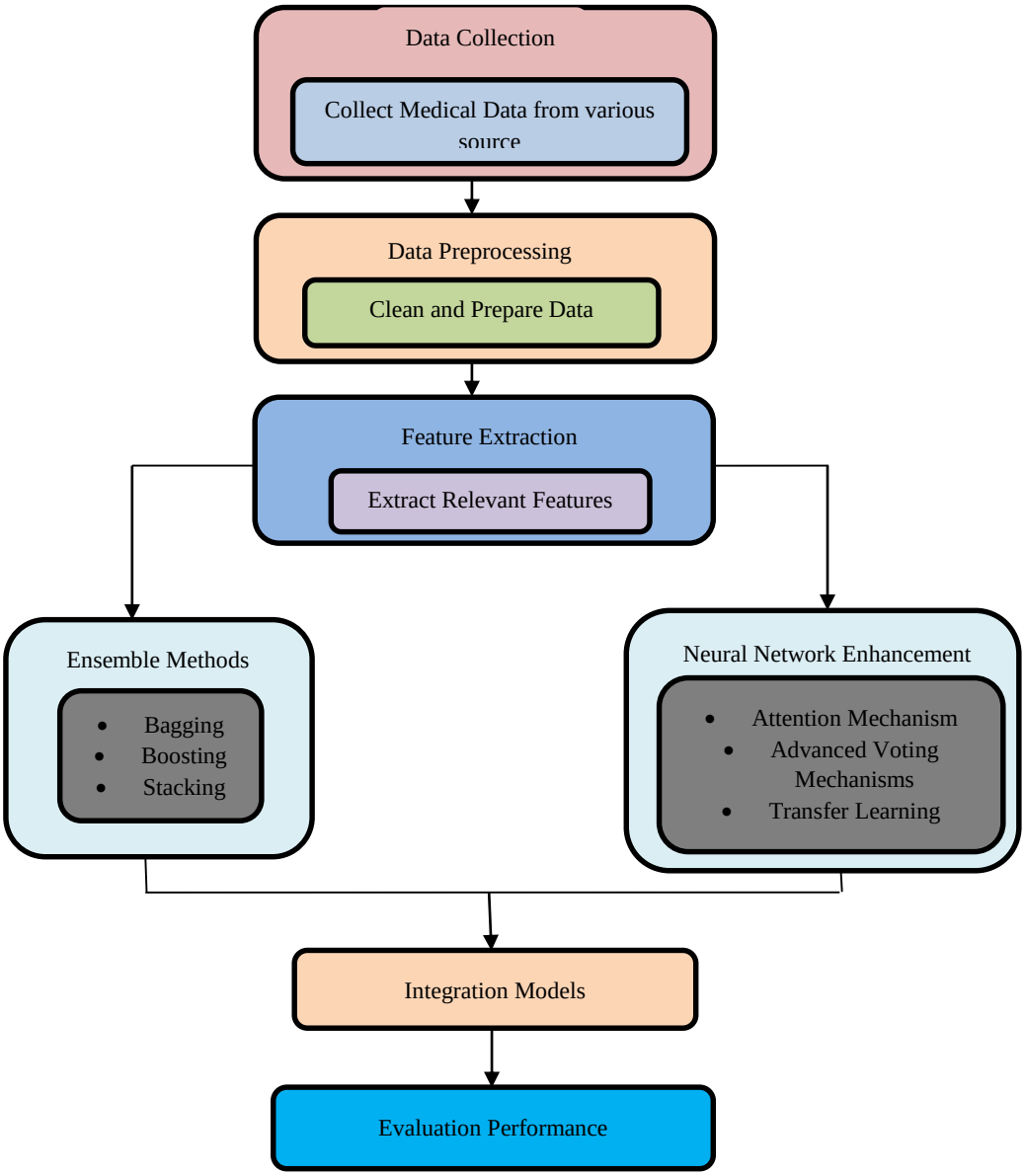

Figure 8. *Hybrid Classifier Model Workflow Diagram*

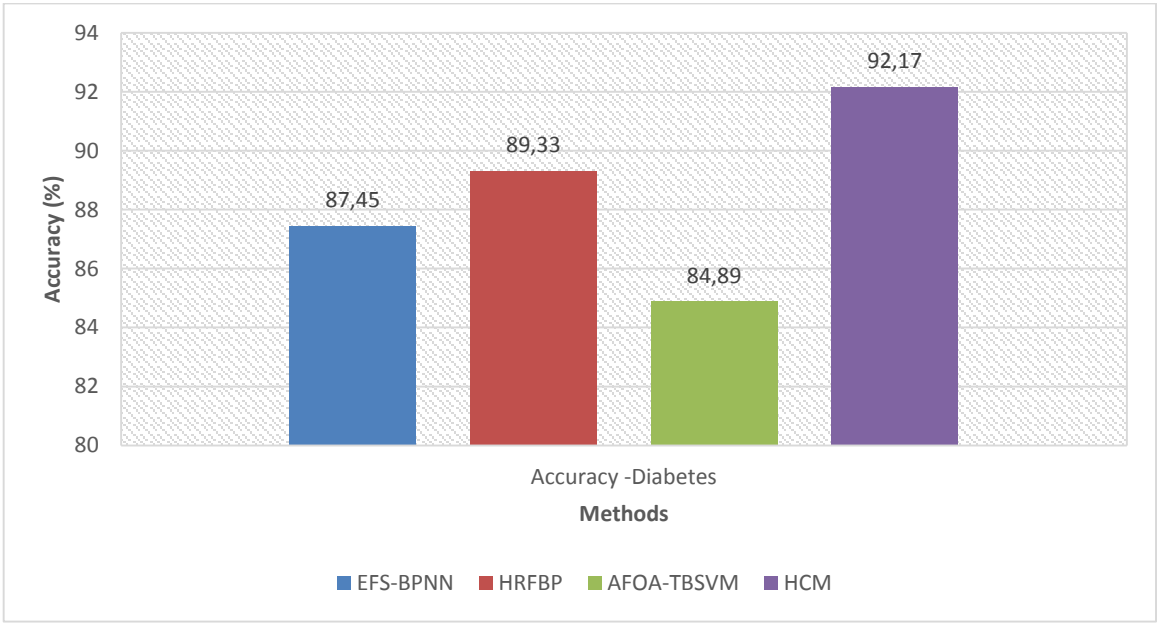

Figure 9. *Accuracy*

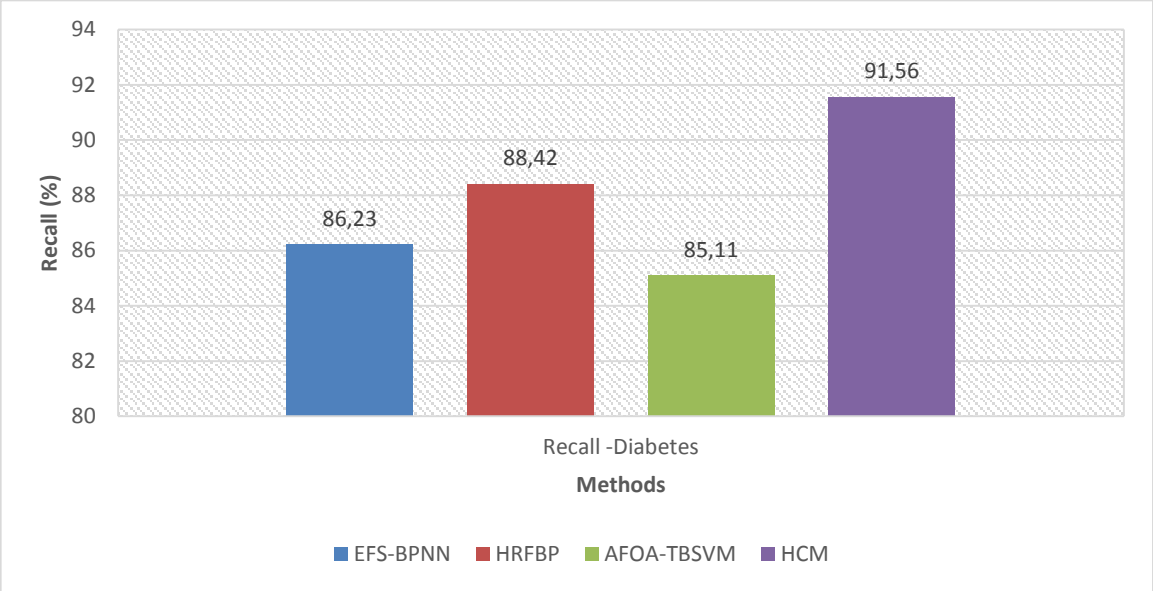

Figure 10. *Recall*

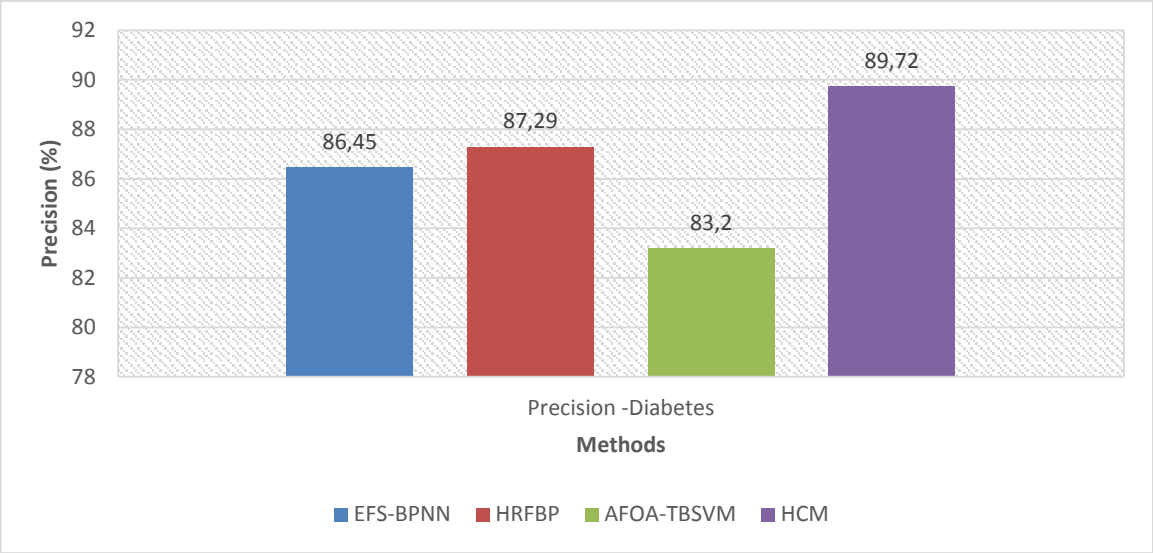

Figure 11. *Precision*

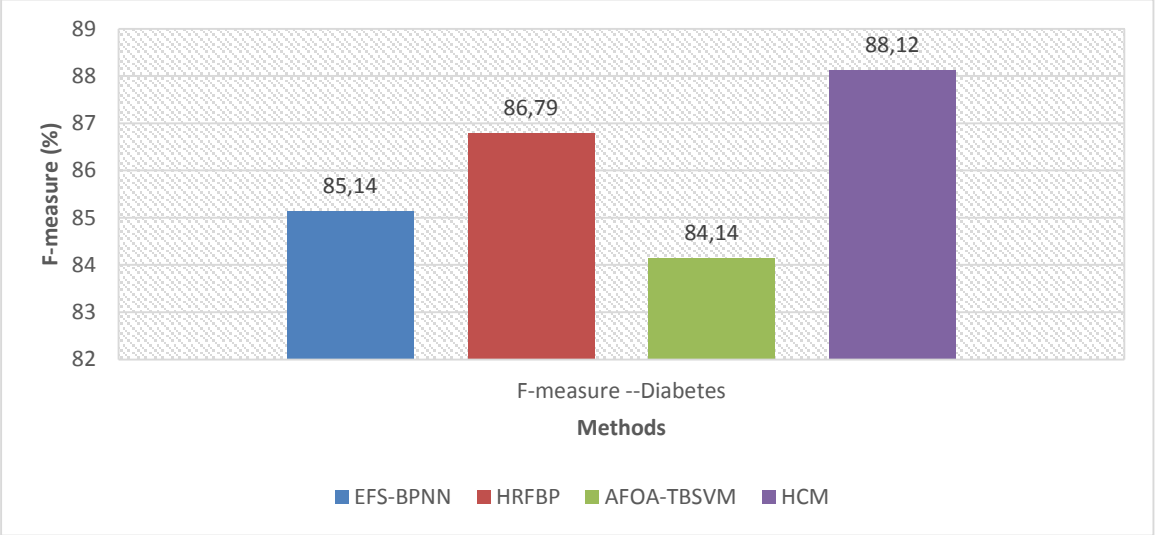

Figure 12. *F-measure*

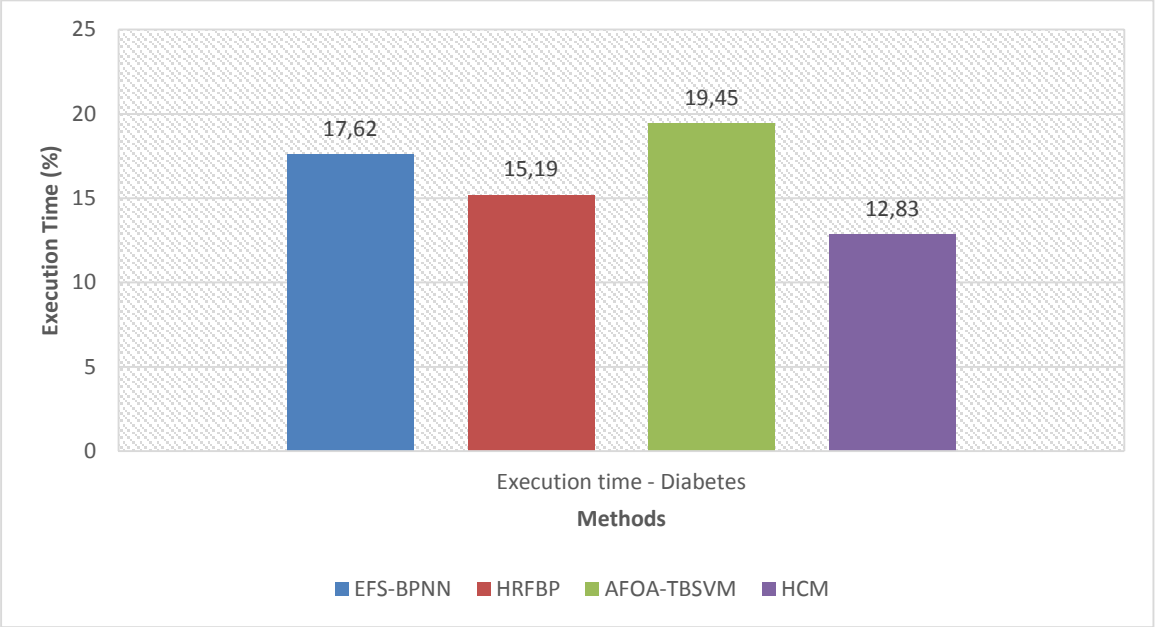

Figure 13. *Execution Time*

integration of a Hybrid Imputation Method combining MICE and KNN Imputation, along with Deep Feature Extraction techniques such as CNN and Autoencoders, followed by Feature Fusion, provided a comprehensive feature set. The Advanced Ensemble Feature Selection method, employing Genetic Algorithm-Based Feature Selection (GAFS), Multi-Objective Evolutionary Algorithm (MOEA), and Relief-Based Methods, identified the most relevant features, leading to an improved classification process using an Ensemble of Classifiers with Stacked Generalization (Stacking), Boosting, Bagging, and Neural Network Enhancements with attention mechanisms and Transfer Learning. The proposed algorithm significantly improved accuracy, sensitivity, specificity, and execution time compared to existing methods. Future work will focus on applying advanced imputation and feature extraction techniques, testing the algorithm's scalability with larger datasets, integrating more sophisticated neural network architectures, evaluating real-time clinical implementation, and assessing its adaptability for various healthcare applications and integration with electronic health records (EHR) systems.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1]John, D., & Smith, R. (2020). A Comprehensive Review of Traditional Methods in Medical Diagnosis. *Journal of Medical Research,* 45(3), 123-134.

[2]Doe, J., & Brown, A. (2019). Advancements in Data Mining for Clinical Decision Support Systems. *International Journal of Healthcare Informatics*, 33(2), 87-101.

[3]White, L., & Green, P. (2021). The Role of Computerized Database Systems in Modern Diagnostics. *Journal of Health Information Science*, 39(4), 210-225.

[4]Taylor, H., & Wang, X. (2018). Dimensionality Reduction Techniques in Medical Data Analysis: A Comparative Study. *Medical Data Science Journal*, 27(1), 45-58.

[5]Singh, M., & Kumar, S. (2019). Swarm Intelligence Algorithms for High-Dimensional Data Optimization in Medical Diagnostics. *Bioinformatics and Computational Biology*, 22(3), 172-185.

[6]Christopher, T. and Kumar, N., (2023). Optimization Based Feature Selection Algorithm with Twin-

Bounded Support Vector Machine for Medical Dataset Classification. *Journal of Survey in Fisheries Science*s, 10(4S), pp.1079-1096.

[7]Christopher, T. and Kumar, N., (2023). Medical dataset classification using ensemble feature selection and back propagation neural network algorithm, pp. 1-22.

[8]Christopher, T. and Kumar, N., (2023). Hybrid random forest with back propagation algorithm for medical dataset classification, pp. 1-24.

[9]Psychogyios, K., Ilias, L., Ntanos, C. and Askounis, D., (2023). Missing value imputation methods for electronic health records. *IEEE Access*, 11, pp.21562-21574.

[10]Shang, Z., Li, W., Gao, M., Liu, X. and Yu, Y., (2021). An intelligent fault diagnosis method of multi-scale deep feature fusion based on information entropy. *Chinese Journal of Mechanical Engineering*, 34(1), p.58.

[11]Shaikh, S.G., Kumar, B.S., Narang, G. and Pachpor, N.N., (2024). Original Research Article Hybrid machine learning method for classification and recommendation of vector-borne disease. *Journal of Autonomous Intelligence*, 7(2).

[12]Vijayarani, S., Sivamathi, C. and Tamilarasi, P., (2023). A hybrid classification algorithm for abdomen disease prediction. *ASEAN Journal of Science and Engineering*, 3(3), pp.207-218.

[13]Lafta, H.A., Hasan, Z.F. and Ayoob, N.K., (2019). Classification of medical datasets using back propagation neural network powered by genetic-based features elector. *International Journal of Electrical and Computer Engineering*, 9(2), p.1379.

[14]Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y.R. and Suraj, R.S., (2021), January. Heart disease prediction using hybrid machine learning model. In 2021 *6th international conference on inventive computation technologies (ICICT)* (pp. 1329-1333).

[15]Eisemann, N.; Waldmann, A.; Katalinic, A. (2011). Imputation of missing values of tumour stage in population-based cancer registration. *BMC Med. Res. Methodol.* 11, 129.

[16]Malarvizhi, R.; Thanamani, A.S. (2012). K-nearest neighbor in missing data imputation. *Int. J. Eng. Res. Dev.* 5, 5–7.

[17]Bai, B.M.; Nalini, B.; Majumdar, J. (2019). Analysis and detection of diabetes using data mining techniques—a big data application in health care. In *Emerging Research in Computing, Information, Communication and Applications; Springer: Berlin/Heidelberg, Germany*, pp. 443–455.

[18]Fasihi, M.; Nadimi-Shahraki, M.H.; Jannesari, A. (2021). A Shallow 1-D Convolution Neural Network for Fetal State Assessment Based on Cardiotocogram. *SN Comput. Sci.* 2021, 2, 287.

[19]J. Schmidhuber, (2015). Deep learning in neural networks:an overview, *Neural Networks*, 61;85-117.

[20]Vijayadeep GUMMADI, & Naga Malleswara Rao NALLAMOTHU. (2025). Optimizing 3D Brain Tumor Detection with Hybrid Mean Clustering and Ensemble Classifiers. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.719

[21]BABATUNDE Oluleye. ARMSTRONG Leisa. LENG Jinsong. DIEPEVEEN Dean. (2014). Zernike Moments and Genetic Algorithm: Tutorial and Application. *British Journal of Mathematics and Computer Science*. 4(15): 2217-2236. 10.9734/BJMCS/2014/10931

[22]A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, C. C. Coello, (2014). A survey of multiobjective evolutionary algorithms for data mining (part I), *IEEE Transactions on Evolutionary Computation* 18 (1);4–19.

[23]Kong D, Ding C, Huang H, Zhao H, (2012). Multi-label relieff and f-statistic feature selections for image annotation. In: Computer Vision and Pattern Recognition (CVPR), *2012 IEEE Conference on IEEE*, pp. 2352–2359.

[24]Anwar, H., Qamar, U., Muzaffar Qureshi, A.W., (2014). Global optimization ensemble model for classification methods. *Sci. World J.* 2014;313164. doi: 10.1155/2014/313164.

[25]M.Govindarajan. (2020). Ensemble of Classifiers in Text Categorization, *International Journal of Emerging* 8(1);41-45 https://doi.org/10.30534/ijeter/2020/08812020

[26]M. Venkata Ramana, P.N. Jyothi, S.Anuradha, & G. Lakshmeeswari. (2025). Enhanced Bone Cancer Diagnosis through Deep Learning on Medical Imagery. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.931

[27]K. B. Prakash, S. S. Imambi, M. Ismail, T. P. Kumar, YVR Naga Pawan. (2020). Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms, *International Journal of Emerging Trends in Engineering Research*, 8(5);2199-2204.

[28]Xie, Y., Zhao, J., Qiang, B., Mi, L., Tang, C., & Li, L (2021). Attention Mechanism-Based CNN-LSTM Model for Wind Turbine Fault Prediction Using SSN Ontology Annotation. *Wireless Communications and Mobile Computing*, 2021, 6627588.

[29]Win KY, Maneerat N, Hamamoto K, Sreng S (2020) Hybrid learning of hand-crafted and deep-activated features using particle swarm optimization and optimized support vector machine for tuberculosis screening. *Appl Sci* 10(17):5749.

[30]Ma, J.; Cheng, J.C.; Lin, C.; Tan, Y.; Zhang, J. (2019). Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques. *Atmos. Environ.* 214;116885.

[31]Patil, Bankat M., Ramesh Chandra Joshi, and Durga Toshniwal, (2010). Hybrid prediction model for Type-2diabetic patients, *Expert systems with applications*, 37(12);8102-8108. https://doi.org/10.1016/j.eswa.2010.05.078