Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 19 No. 13 (2023) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v19i13.40145

PAPER

Evolutionary Optimization Algorithm for Classification of Microarray Datasets with Mayfly and Whale Survival

Peddarapu Ramakrishna(✉), Pothuraju Rajarajeswari

Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India

peddarapuramakrishna2022@gmail.com

ABSTRACT

In the field of bioinformatics, a vast amount of biological data has been generated thanks to the digitalization of high-throughput devices at a reduced cost. Managing such large datasets has become a challenging task for identifying disease-causing genes. Microarray technology enables the simultaneous monitoring of gene expression levels, thereby improving disease diagnosis accuracy for conditions like diabetes, hepatitis, and cancer. As these complex datasets become more accessible, innovative data analytics approaches are necessary to extract meaningful knowledge. Machine learning and data mining techniques can be employed to leverage big and heterogeneous data sources, facilitating biomedical research and healthcare delivery. Data mining has emerged as a vital tool in the medical field, providing insights into illnesses and treatments and enhancing the efficiency of healthcare systems. This thesis aims to present a novel hybrid technique for feature selection using amalgamation wrappers. The proposed approach combines the Mayfly and whale survival strategies, leveraging the strengths of both algorithms. The model was evaluated using various datasets and assessment criteria, including precision, accuracy, recall, F1-score, and specificity. The simulation results demonstrated that the proposed integrated optimization model exhibits improved classification performance with 12% higher accuracy in disease diagnosis.

KEYWORDS

microarray technology, machine learning, data mining, optimization, mayfly, whale survival techniques

1 INTRODUCTION

Bioinformatics data classification is a key machine learning job that has been widely adopted for high-dimensional microarray datasets. It is critical to determine the most important characteristics for improving classification accuracy since classification performance is heavily dependent on the quality of the learning algorithm [1]. Due to the increasing amount of information that must be processed, there is

Ramakrishna, P., Rajarajeswari, P. (2023). Evolutionary Optimization Algorithm for Classification of Microarray Datasets with Mayfly and Whale Survival. *International Journal of Online and Biomedical Engineering (iJOE)*, 19(13), pp. 17–37. https://doi.org/10.3991/ijoe.v19i13.40145

Article submitted 2023-04-04. Revision uploaded 2023-07-10. Final acceptance 2023-07-10.

some duplication and uninteresting characteristics, resulting in excessive training as well as classification time [2].

Feature selection is a typical strategy used to handle the exponential rise of features in high-dimensional data to reduce information in order to enhance classification performance. The methods for feature selection may be divided into three types: wrapper, filter, and embedding [3]. In the absence of different classifiers, characteristics in filter approaches are categorized based on intrinsic information [4]. Each intrinsic quality evaluates and ranks features using ranking algorithms such as weights, dependence, and distance measures. Such ranking algorithms are unquestionably advantageous when dealing with huge amounts of dimensional information [5].

Generally, gene expression data comprises a significant number of genes, necessitating the use of analytic tools in order to get relevant information [6]. Because of the huge quantity of gene expression knowledge (features or genes) that can be utilized to detect common patterns within a collection of samples, the development of gene expression technology has rendered microarray data increasingly useful in cancer research categorization. Microarrays are a popular tool for detecting cancer cells by examining deoxyribonucleic acid (DNA) proteins for additional gene research. The gene expression matrix is an array that organizes microarray data, with each row representing a single gene and each column representing an experiment condition [7]. Microarray technology can provide valuable insights into disease-gene correlations.

Deoxyribonucleic acid has long been thought to be the genetic material in living systems [9]. DNA is constructed from four distinct deoxynucleotide monomers as a generic substance with sequence programmability. Each monomer is made up of deoxyribose, a phosphate group, and any 1 of 4 nitrogen-containing nitrogenous bases (cytosine [C], guanine [G], adenine [A], or thymine [T]) (respectively marked in green, blue, and orange colors in Figure 1). A phosphodiester link is formed between the deoxyribose of one monomer and the phosphate molecule of the next monomer. A single-stranded DNA (ssDNA) chain is formed by the orderly and continuous linking of a given number of monomers via phosphodiester bonds [10].

However, in this hybrid optimization mechanism, an amalgamation of MayFly and whale survival techniques is proposed, taking into consideration the benefits of both algorithms. The contribution of the paper is as follows:

- To choose the most informative and significant genes for the classification problem.
- To reduce data and scale down the storage requirements for improved performance.
- Generating a simpler model that allows for greater speed and simplicity.

The rest of the paper is organized as follows: Section 2 describes related works. The proposed methodology is described in Section 3. Section 4 discusses the performance evaluation. Finally, the conclusion and future works are presented in Section 5.

2 RELATED WORKS

To boost the overall effectiveness of the classification model, Sahu and Shrivas (2022) presented a genetic search with the Wrapper Subset Evaluator approach for feature selection [11]. They also classified CKD and non-CKD data using the Bayes Network, Classification and Regression Tree (CART), and J48 classifier. The genetic algorithm chooses the best features from the CKD dataset and compares classifier

performance to other genetic search FSTs. All classification models perform better with GSBFST than without FST or current genetic search FSTs.

Ibraheem et al. (2022) demonstrated the application of a multi-objective iterative method for feature subset selection in computerized BC diagnosis [12]. The logistic regression (LR) approach was used in this paper for the classification job. For subset feature selection, the projected framework incorporates the TLBO and LR. The assessment of an event's occurrence probability using LR-based categorization is determined by the similarity of the provided data points. According to the findings, the projected technique generated higher classification accuracy for the BC dataset.

To anticipate breast cancer, Farid et al. (2021) suggested a Composite Hybrid Feature Selection Approach Focused on Optimizing the Genetic Algorithm (CHFS-BOGA). This hybrid feature selection strategy combines the benefits of three filter feature selection techniques with an OGA to choose the best features and increase classification process efficiency and scalability. OGA is proposed by enhancing the initial population generation as well as genetic operators utilizing filter technique outcomes, as some previous knowledge suggests. The results reveal that for optimum feature selection, the hybrid feature selection strategy outperforms single filter approaches with principal component analysis (PCA) [13].

Alelyani (2022) presented a bagging-based ensemble strategy to increase feature selection consistency in medical datasets through data variance reduction [14]. An experiment is carried out utilizing four real-world medical datasets, each of which has high dimensionality and a small sample size. In this technique, the bagging technique is utilized to decrease data volatility, which increases the reliability of the process of selecting features. The suggested method significantly improves selection stability while retaining classification accuracy. This is accompanied by a rise in classification accuracy in the majority of cases, indicating the stated outcome of stability.

Sangaiah and Kumar (2021) introduced a novel hybrid breast cancer prediction algorithm [15]. For breast cancer diagnosis, the expected method employs relief attribute minimization with an entropy-based genetic algorithm. To handle datasets with high dimensionality and uncertainty, a hybrid mix of these approaches is utilized. The data were acquired from the Wisconsin breast cancer database and were classified based on several features. This method's performance is tested, and the findings are compared to other feature selection approaches in the literature.

To overcome optimization challenges, Zainudin et al. (2021) integrated relief-f with the differential evolution (DE) attribute selection approach. In this study, population numbers as well as generation size were calculated adaptively using the number of characteristics from relief-f. Using 10 datasets from the UCI machine learning repository, the effectiveness of this method is compared to that of different feature selection strategies in order to demonstrate its superiority [16].

Shankar et al. (2020) suggested a system for classifying CKD that uses an inspired optimization model and a learning procedure. This approach uses the Ant Lion Optimization (ALO) strategy to pick appropriate aspects of renal data for the classification procedure. The CKD data is then sorted using a deep neural network (DNN) depending on the specified characteristics. When compared to previous data mining classifiers, performance comparison shows that this model achieves greater classification accuracy, precision, F-measure, and sensitivity measures [17].

Diabetes is predicted by Alam et al. (2019), utilizing important factors, and the link between the various qualities is also described. For diabetes, many techniques are utilized to assess relevant attribute selection as well as clustering and prediction, including association rule mining. The principal component evaluation approach was used to pick significant features. The data show a substantial relationship

between diabetes, body mass index (BMI), and glucose levels, as determined by the Apriori approach. Diabetes was predicted using artificial neural networks (ANN), RF, and K-means clustering approaches.

Recent advanced feature selection approaches use the power of optimization algorithms to choose a subset of important characteristics to produce better classification results in order to improve the performance of feature selection methods. The following research papers provide an overview of optimization-based feature selection approaches used in medical illness detection utilizing a hybrid optimization mechanism, which is a combination of MayFly and whale survival strategies.

3 PROPOSED METHODOLOGY

Microarrays are used to diagnose diseases such as hepatitis, diabetes, and breast cancer. After loading a dataset, preprocessing approaches were employed to exclude and substitute missing-value features. The training group was then separated into two sub-data groups: training samples and testing samples. The training sub-data were used for building classifiers and evaluating individuals throughout the evolutionary processes, while the test sub-data were utilized for evaluating the repository's results. Preprocessed data is sent into the feature selection stage, which employs a hybridization of the MayFly and whale survival algorithms. As a result, the chosen features are subjected to a hybrid of the convolution neural network (CNN) and the HopeField Classifier. Figure 1 shows a block diagram for microarray dataset classification.

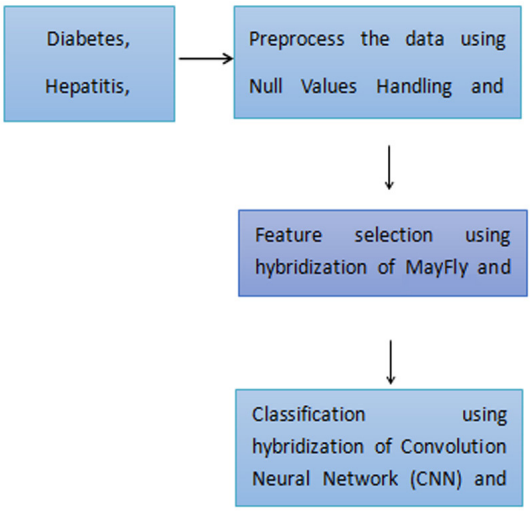

Fig. 1. Block diagram for micro array dataset classification

3.1 Dataset description

Diabetes dataset: We obtained twelve microarray datasets from GEO related to diverse case-control diabetes investigations. Those datasets, comprising samples from several disease phenotypes were further subdivided, and the resultant 20 subsets were categorized tissue-wise as follows: skeletal (eleven sets), subcutaneous adipose (four sets), peripheral (three sets), and liver (two sets). The datasets and sub-divisions were chosen based on prior research on candidate gene identification

in type 2 diabetes mellitus [19]. Human patients were also followed in certain research studies for additional diabetes-related clinical characteristics such as illness, family history, food, and physical training regimen. As a result of the inclusion of many environmental factors that are critical for the clarification of diabetes pathogenesis, these datasets may be regarded as an outstanding collection of T2D gene-expression profiles.

Patients hospitalized at the Liver Department of the Hospital Clinic in Barcelona (2007–2009) exhibiting clinical, analytical, or histological signs of AH were prospectively incorporated into the hepatitis dataset [20]. The inclusion requirements have already been defined. 13 24 25 All of the patients had an AH histological diagnosis. The research excluded patients with cancer or any other probable cause of liver damage. A transjugular technique was used to acquire liver samples. We used individuals with chronic hepatitis C-induced liver problems (HCV) as controls (n¼18). All of the individuals had HCV genotype 1 and had not previously received antiviral therapy. According to Kleiner's criteria, we included a cohort of individuals with morbid obesity and concomitant nonalcoholic steatohepatitis (NASH) (n¼20). During bariatric surgery, these individuals had a laparoscopic liver biopsy. An experienced liver pathologist examined all patients' liver specimens, and a portion of the biopsy was immersed in an RNA-stabilizing solution (RNAlater, Ambion, Austin, Texas, USA). The procedure followed the ethical principles of the 1975 Helsinki Declaration and was accepted by the Ethics Committee of the Hospital Clinic of Barcelona; only patients with written informed consent were enrolled.

Breast cancer dataset: The Gene Expression Omnibus dataset was used to acquire breast cancer datasets. The criteria were "Organism: Homo sapiens" and "Experiment Type: Transcript Profiling by Array." The platform Affymetrix Gene Chip Human Genome U133 Plus 2.0 Array (CDF: Hs133P Hs ENST, version 10) was used (Affymetrix, Inc., Santa Clara, CA, 95051, USA). All datasets contained the GEO authorization code, platforms, sampling procedure, quantity of sample, and gene regulation data. The microarray platforms, as well as the hgu133plus2 annotating platform of probes, were used to determine the proteins with differential expression. The R programming language was used in addition to the Bio-conductor packages Biobase, hgu133a2cdf, Affy, Limma, AffyRNADegradation, AnnotationDbi, hgu133plus2cdf, Annotate, and AffyQCReport.

3.2 Preprocessing of data

The handling of undefined values is the first portion of the preprocessing layer. It should be noted that such actions are often represented by NULL values; nevertheless, they can have specific attribute connotations in the temporal system. Furthermore, it has substantial speed constraints, as NULL values are not indexed at all using Btree index structures. Because its main limitation is the limited cardinality of column values, the bitmap index data structure, as the second index type based on quantity, somehow doesn't provide sufficient power, even though it can handle NULL values; however, for date and sensor data processing, such an approach is totally unsuitable. Bitmap indexes are most commonly employed in data warehouses and decision-support systems. Furthermore, when numerous update statements are utilized, performance suffers dramatically. It is important to stress that temporal information is primarily distinguished by rapid update streams. Depending on the characteristics of attribute X and the technique utilized, missing data pieces can be handled in a range of ways. If X is an integer, for

illustration, missing data is commonly "filled" by averaging X or estimating gave attention to other independent features. Assuming X is a low-cardinality category attribute encoded with one-to-m bit characters, an m-dimensional array of 0 might represent the missed case. Moreover, several machine-learning algorithms, such as Naive Bayes and decision trees, completely disregard or accept incomplete information as an additional value. The suggested processing approach treats missing data similarly to any other valuation. This may be accomplished by adding a new value for X, a null-value X_O with the probability of the target, X = X_O, utilizing the standard formula:

$$S_0 = \tau(n_0)\frac{n_{0_\gamma}}{n_0} + (1-\tau(n_0))\frac{n_\tau}{n_{TR}} \quad (1)$$

If the occurrence of a missing parameter for feature X has predictive consequences for the accused, S O will record that data. If the missing data are unrelated to the defendant, S O will converge to the target's posterior distribution, corresponding to a "neutral" portrayal of the random variable. In the sklearn-pandas package, category imputer is a unique method for working with categories with missing information. It is used on information columns in the category "string," so it substitutes null values with the column's most frequently occurring value. Because the scikit-learn modules' imputing algorithms are limited to numerical data, researchers who use them cannot impute missing category values. As a consequence, although the categorical imputer approach is beneficial for imputing missing category values, the imputing techniques of the scikit-learn module may be used for numerical data. Consider probability estimate formula (3), which also works for continuous targets with a category feature such as:

$$S_i^5 = \tau(n_i)\frac{n_{i1}}{n_i} + (1-\tau(n_i))\frac{n_\tau}{n_{TR}} \quad (2)$$

The calculation involves determining the target possibility for a cell value by combining the frequency-based goal probabilities in the cell with the posterior distribution n/n TR. Instead of using the target's nTR prior probability as the "null hypothesis," it is fairer to use the projected possibility at the subsequent highest aggregation level inside the characteristic hierarchy.

$$S_i^5 = \tau(n_i)\frac{n_{i1}}{n_i} + (1-\tau(n_i))S_i^4 \quad (3)$$

It should be obvious how this formula automatically modifies the assumption based on the concentration of the data throughout the hierarchy's various levels. Every other numeric value in the dataset is recognized with its own set of rules, calculated, and replaced with the appropriate language label. This procedure is done for each numerical score in the provided dataset. This entire approach is automated in order to efficiently preprocess and prescribe the dataset.

3.3 Feature selection using hybrid Mayfly and whale optimization

Many species in nature exhibit similar foraging behaviors. Whales, for example, exhibit a distinct predatory behavior in foraging known as bubble net foraging.

The WOA simulates whale predation behavior by designing a smaller encircling mechanism and an upward spiral assault path. When hunting for food, mayflies travel from one location to another in search of ample food. They would hunt prey in spiral form in the air whenever rich prey was located. This work offers a hybrid Mayfly and whale optimization algorithm (MWOA) for complex optimization problems that combines the WOA's shrinking encircling mechanism with the MOA's mating behaviors, considerably increasing the algorithm's local and worldwide searching abilities. Levy flying is a method that controls local search through random walking behaviors.

The seagull optimization algorithm, on the other hand, will converge prematurely. This research explores including the levy flight technique into the contraction-encircling mechanisms of the WOA as well as the local pairing of the MOA, which increases the exploiting ability and prevents the algorithm's early convergence.

Exploitation phases. Equations (1) and (2) mimic the circular habits of humpback whales. The leading solution is considered the target prey by the algorithm, with some other alternatives seeking to close in on the targeted prey.

$$\begin{aligned} u' &= \left|\beta' . X'_b(t) - X'(t)\right| \\ X'(t+1) &= \left|X'_b(t) - \propto' . u'\right| \end{aligned} \tag{4}$$

Here, 't' is the current iteration, X'_b is the best-so-far solution, $\propto'$ and β' are coefficient vectors that are shown as follows:

$$\begin{aligned} \propto' &= 2a' . r' - a' \\ \beta' &= 2 . r' \end{aligned} \tag{5}$$

Here, a' is a linear decreasing coefficient that decreases from 2 to 0 throughout the iteration process, and r' is a random vector in the iterative procedure (0, 1). The values of $\propto'$ and β' vectors are adjusted to control the different locations of the present position relative to the best-so-far solution. In (2), the algorithm considers that the prey is the best-so-far answer, changes the humpback whale's current location to a position near the prey, and simulates the circumstance of an encircling prey. Two mathematical models are developed to imitate the bubble-net attack on humpback whales.

Shrinkage circling method: This model is accomplished by lowering the value of the vector' linearly. The fluctuation band of the coefficients vector $\propto'$ is among $(-\propto', \propto')$ depending just on vectors a and randomized vector r', where a' is lowered from two to zero during the iteration process.

The model first estimates the range from it to the target, before the humpback whale surrounds the prey in a logarithmic circular movement, as depicted in the mathematical formula below:

$$X'(t+1) = U' \times e^{r\varphi} \times \cos(2\pi\varphi) + X'_b(t) \tag{6}$$

Here $U' = \left|X'_b(t) - X'(t)\right|$ is a parameter for defining the form of the exponential spirals, which is a random number in (1, 1).

During the optimization procedure, the humpback whale will deep dive and then begin to emit a spiral bubbles surrounding the prey as it moves upward toward the surface. In a spiral, the humpback whale progressively retreats inside the ring while chasing food. According to the hunting behavior, the diminishing circle, and the

spiral-shaped pathway each have the same implementation chance for revising the humpback whale's location throughout the illustrations.

Movements of male mayflies. During iterations, male mayflies in a swarm will continue the exploration or exploitation operation. The velocity will be changed depending on their current fittest f (x i) and the fittest values in previous paths $f(x_{hi})$. If $f(x_i) > f(x_{hi})$, the male mayflies will then change their velocity based on their current velocity, the space between themselves and the global optimal location, and the historic best trajectory will be:

$$v_i(t+1) = g.v_i(t) + a_1 e^{-\beta r_p^2}\left[x_{hi} - x_i(t)\right] + a_2 e^{-\beta r_g^2}\left[x_g - x_i(t)\right] \tag{7}$$

In this case, g is a variable that decreases linearly from its maximum to a lesser number. The constants a_1, a_2, and β are used to equalize the quantities. The parameters r_p^2 and r_g^2 are employed to calculate the Cartesian distances among individuals and their historical best position, as well as their global position in the swarm. The Cartesian range would serve as the length array's second norm.

$$\left\|x_i - x_j\right\| = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})} \tag{8}$$

On the other hand, if $f(x_i) < f(x_{hi})$, then the male mayflies would update their velocity from the current with a random dance coefficient d:

$$v_i(t+1) = g.v_i(t) + d.r_1 \tag{9}$$

Here, r_1 is a uniformly distributed random value drawn from range [–1, 1].

Movements of female mayflies. Female flies have the ability to adjust their velocity in an unusual way. Female mayflies have wings that last only between 1–7 days, so they might be in a rush to locate male mayflies to breed with and reproduce themselves. As an outcome, their velocity would change depending on which male mayfly she wanted to engage with. The MO technique calculates the finest female and male mayflies to be the initial partners, followed by the second best female and male mayflies, and so on. As an outcome, if I is the i-th female mayfly, $f(y_i) < f(x_i)$

$$v_i(t+1) = g.v_i(t) + a_3 e^{-\beta r_{mf}^2}\left[x_i(t) - y_i(t)\right] \tag{10}$$

Here, a_3 is another constant. r_{mf}^2 gives Cartesian distance. On the contrary, if $f(y_i) < f(x_i)$ Female flies will refresh its present velocity with some other randomized dance. fl:

$$v_i(t) = g.v_i(t) + fl.r_2 \tag{11}$$

Where, r_2 is random number within uniform-distribution in range [–1, 1].

Exploration phase. The technique causes the answer to be distant the best current solutions during the exploration phase and randomly traverses the search space. As a result, the WOA algorithm picks the reference solution at random using a random number of the $\propto'$ vector larger than + 1 or even less than 1. This method, in conjunction with $\propto' > 1$, enables the algorithm to do global exploration. Here is the mathematical model:

$$U' = \beta'.X'_{rand} - X'$$
$$X'(t+1) = X'_{rand} - \propto'.U' \quad (12)$$

Here X'_{rand} is the location vector of a solution picked at random out from present populations

Mating of mayflies. Every one of the leading 50% females and males would be mated then given a pair of progenies. Their kids would be created at random from their parents:

$$offspring\,1 = L*male + (1-l)*female$$
$$offspring\,2 = L*female + (1-l)*male$$

Here, *L* gives random no, in the Gauss distribution.

As the gap between *j*th and *i*th persons expands, *r j* will seem to be larger. The *i*th individual may be referenced to the global optimum, the *j*th individual's historic ideal situation, or its mate. Therefore, due to the declination of the negative exponential function, the weights for the range would be lowered. This means that when the distance between *p j* and *p I* grow, so do the weights and the overlayed velocities *v p*. If the difference between *p j* and *p I* is narrowed, the weight will be increased instead. As a consequence, when *p j* is sufficiently separated from *p I*, it adjusts its speed with a lesser range, and when *p j* is sufficiently close to *p I*, it changes its speed with a larger amplitude. This also means that when they are far away, they approach each other more slowly; conversely, whenever they come face-to-face, they slide away more quickly. These possibilities are just unacceptable.

Classification using hybridization of convolution neural network and Hopefield (CNN_HF). The Hopfield architecture is a completely connected neural network that can recover previously stored memories from loud and distorted inputs. The Hopfield network is made up of *N* neurons linked together by symmetrical bidirectional linkages. The binary variable will be used to represent the activity of neuron i. *x*_i∈ {+ 1, –1} corresponds to the two conceivable states of neurons: firing (+1) or silence (–1). As a result, the network's state may be depicted as a binary array *x* = (*x* 1,...*x N*), with the ith component, *x i*, representing the state of neuron *i*. The interactions of neurons are represented in the connecting matrix, which is a *NN* real square matrix with self-interaction terms, with entries *w j* defining the strength of the link between neurons. *i* and *j*. Where x_i^t is the temporal development of the network specified by the following updated rule based on the state of neuron I at time t:

$$x_i^{t+1} = sign\left[\sum_{j=1}^{N} w_{ij} x_i^t\right] \quad (13)$$

(1) describes dynamics that may be done either asynchronously or synchronously. In the first scenario, the status of all neurons is concurrently updated at time *t*. Asynchronous updating, on the other hand, modifies the state of one node at a time based on the status of its neighbors. We picked asynchronous updating because it has superior convergence features since it reduces misleading cycles. In general, the Hopfield system is utilized for pattern storage and recovery. A collection of *p* patterns must be kept in the network, $\delta^{\mu} = (\delta_1^{\mu}, \ldots . \delta_N^{\mu})$ with $\mu = 1, \ldots, p$

should be stable fixed points in the dynamics. This may be accomplished in two steps. First, it can be demonstrated that an asynchronous update as well as a symmetrical non-negative vector W are necessary criteria for the recursion to reach a stable state. The energy function is commonly used to demonstrate the truth of this assertion:

$$E = -\frac{1}{2}\sum_{i,j=1}^{N} w_{ij} x_i x_j \tag{14}$$

Demonstrating a Lyapunov function for something like the system If this is confirmed, energy doesn't grow at every state transition; and system development results in a local energy minimum. Secondly, it can be demonstrated that with the draw the conclusion matrix selection

$$w_{ij} = \frac{1}{N}\sum_{\mu=1}^{P} \delta_i^{\mu} \delta_j^{\mu} \tag{15}$$

called the Hebb rule, when p isn't large, patterns $\delta^{\mu} = (\delta_1^{\mu}, \ldots \delta_N^{\mu})\ \mu = 1, \ldots p$. With this rule, the number p of patterns that may be stored in the network is finite and proportionate to the number of neurons N. There is a crucial variable p c, termed storage capacity, and only if the number of patterns is less than pc can the model retrieve them. We impose a cutoff on the signature size to counteract the occurrence of imbalanced microarrays, which may cause a bias in favor of bigger microarrays. As a result, if the size N of the signature connected with the class is greater than 1, genes are chosen at random first from the relevant signature. When $N\mu > \Gamma$ all neurons corresponding to genes eliminated from the signature are set to 0 in the input configuration. As a result of the updating rule, they play no further part in the model's temporal development. We paired the signal reduction with something like a resampling approach to avoid information loss. As a result, the threshold is used M1 times (the default M = 100). The class to which the Hopfield models conformed in the majority of the M trials is chosen as the final outcome of our study for each sample.

$$p_{error} = \frac{1}{2}[1 - \mathrm{erf}\left[\sqrt{\frac{N}{2p}}\right]] \tag{16}$$

where $N = N(\Gamma)$ is used for the program as well as in the following. Assuming that a pattern is made up of N bits, the probability of an error-free retrieval of a recorded pattern is (1-p error) N, which must be larger than some fixed number, such as 0.99. Given that the factorial expansions as well as the smallest terms in the p error should be minimal, we obtain

$$p_{error} = \frac{1}{2}[1 - \mathrm{erf}\left[\sqrt{\frac{N}{2p}}\right]] < \frac{0.001}{N} \tag{17}$$

Here, $p << N$ and $erf(x) = 1 - \mathrm{x}$

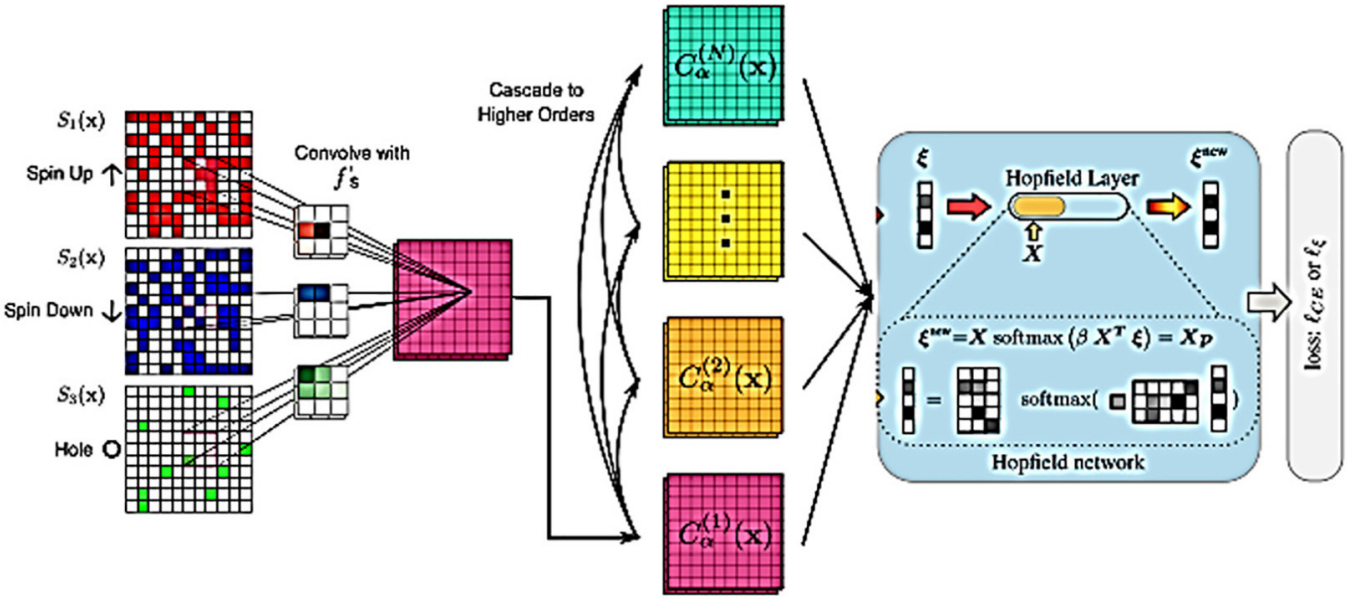

Fig. 2. Convolution neural network and Hopefield (CNN_HF) based classification

Therefore, it must be minimal in order to reduce noise and have balanced signatures, but it must also allow N to meet equation 6. Given that normally 2p6, the threshold value of N () is between 70 and 100, according to equation 6, because the configurations we are keeping in the networks are correlated, we set a number such that the value of N is much greater than the aforementioned threshold and chose = 200 as the default value for our analysis. To learn how to discern between these 2 ideas, we propose a neural network architecture (CNN HF), which is a hybridization of convolution neural network and Hopefield, as given in Figure 2.

The input to the network is an image map having three channel $\{S_k(x)[k = 1,2,3]\}$, here $S_1(x) = n \uparrow (x)$, $S_2(x) = n \downarrow (x)$, $S_3(x) = nhole\ (x)$ Because the models we are considering are limited to one Hilbert space, each input can only take on the values 0 or 1. The CCNN creates nonlinear "correlation maps" from this input, which contain data on local spin-hole interactions up to a certain order N across the snap. This procedure is parameterized by f (a, k(= 1,... M)) learnable 3-channel filters, wherein M represents the number of filters in the models. The mappings for the specified filter are defined as follows:

$$\begin{aligned} C_\alpha^{(1)}(x) &= \sum_{a,k} f_{a,k}(a) S_k(x+1) \\ C_\alpha^{(2)}(x) &= \sum_{(a,k)=(b,k)} f_{a,k}(a) f_{a,k'}(b) S_k(x+a) S_k(x+b) \\ C_\alpha^{(N)}(x) &= \sum_{(a,k1)=(b,k1)} f_{a,k}(a) f_{a,k'}(b) S_k(x+aj) \end{aligned} \tag{18}$$

In this case, it runs over the filter's convolutional window. Traditional CNN uses just one of these procedures, which are alternated with a nonlinear activation function like h RELU $(x) = \max(0, x)$. The problem with common non – linear functions is that typically combine all levels of correlation into the extracted feature, making it impossible to determine what traditional networks are measuring. Each level of our nonlinear convolutions, on the other hand, $C_\alpha^{(N)}(x)$ is especially created to understand n-site semi-local correlations at site x that manifest as patterns in the

convolution layers f. It is necessary to do a direct calculation of nonlinear convolution layers up to order N, $O((KP)^n$ per site; here P gives the pixel count and K gives the number, of microarrays.

4 PERFORMANCE ANALYSIS

The experimental results are analyzed using Python software and the parameters precision, accuracy, specificity, F1-score, and recall. The comparison is made for three different datasets, including hepatitis, diabetes, and breast cancer. The following are the performance metrics:

Accuracy: To compare the total expected values to the number of predictors for successfully categorized occurrences. It is expressed mathematically:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

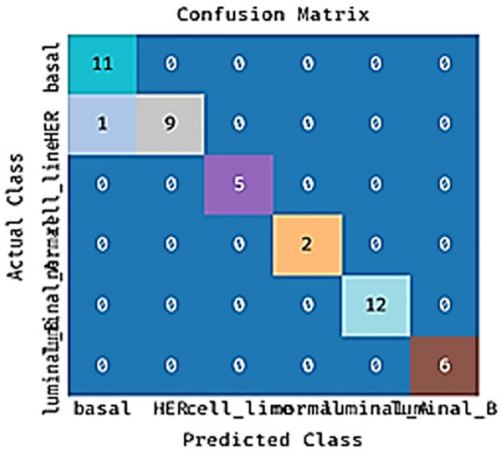

Fig. 3. Confusion matrix for breast cancer testing validation

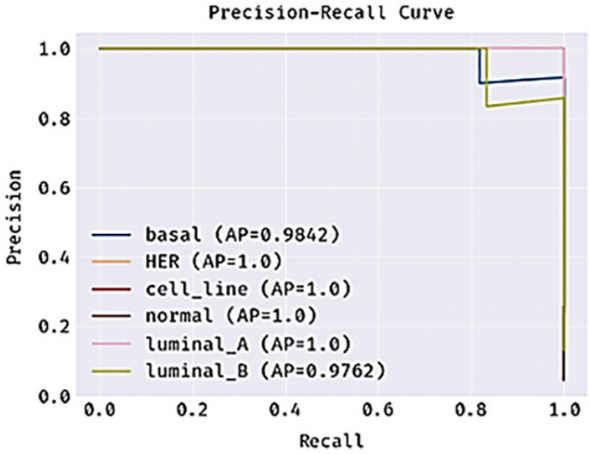

Fig. 4. Precision-recall curve for breast cancer testing validation

Recall or sensitivity: Recall provides a ratio of the value of the correct prediction to the total prediction values. It is defined in the equation.

$$Recall = \frac{TP}{TP + FN}$$

Precision: Precision is expressed in terms of true positives to total anticipated values. It is expressed mathematically.

$$\text{Precision} = \frac{TP}{TP + FP}$$

F1 – Score: It is involved in the computation of the ratio of the average value of precision and recall. F1-Score expressed mathematically:

$$\text{F1} - \text{Score} = 2 * \frac{Precision * recall}{precision + recall}$$

Figure 3 depicts the discriminant function for breast cancer test validation; here, rows indicate the anticipated class, and columns give the actual data class.

The diagonally colored cells represent the tested systems that are categorized properly or inaccurately. The column here on the right side represents each anticipated class, whereas the row at the bottom reflects each actual class's performance. Table 1 shows training and testing validation for the breast cancer dataset:

Table 1. Train and test validation for breast cancer dataset

Parameter	Testing Values	Training Values
accuracy	0.9783	0.9783
Precision	0.9861	0.9861
Recall	0.9833	0.9833
Specificity	0.9952	0.9952
F1-score	0.984	0.984

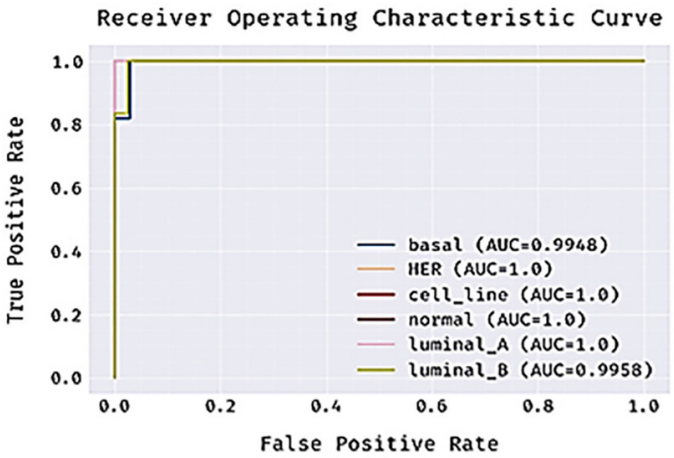

Fig. 5. ROC curve for breast testing validation

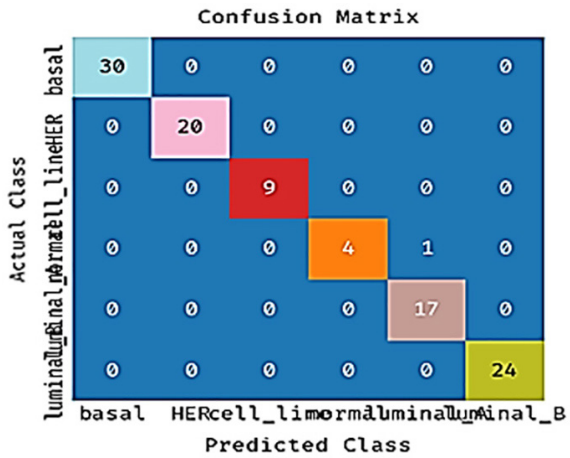

Fig. 6. Confusion matrix for breast cancer training validation

Figure 4 depicts the precision-recall curves for breast cancer test validation, where the x-axis represents recall and the y-axis represents accuracy. Average accuracy (AP) is 0.9842 for basal 1.0 with HER, cell line, normal, luminal A, and luminal B, and 0.9762 for luminal B. Figure 5 depicts the ROC curve for breast cancer testing verification, where the x-axis represents the false positive rate and the y-axis represents the TRUE POSITIVE RATE parameter. The AUC for HER, cell line, normal, luminal A, and luminal B is 0.9948 with basal 1.0 and 0.9958 for luminal B.

Figure 6 depicts the scatterplot for prostate cancer training validation, where the rows indicate the predicted class and the columns represent the actual data class. The electric networks that are successfully and erroneously categorized are shown by diagonally colored cells. The column just on the right side represents each anticipated class, whereas the row at the bottom reflects each actual class's performance.

Figure 7 depicts the precision-recall curves for breast cancer train validation, where the x-axis represents recall and the y-axis represents accuracy. AP is 0.9989 for baseline, 1.0 for HER, cell line, normal, luminal A, and luminal B, and 0.9762 for luminal B. Figure 8 depicts the ROC curve during breast cancer screening validation, where the x-axis represents the false positive rate and the y-axis represents the true positive rate parameter. The AUC is 0.9948 for basal, 1.0 for HER, cell line, normal, 0.9967 for luminal A, and 0.9992 for luminal B.

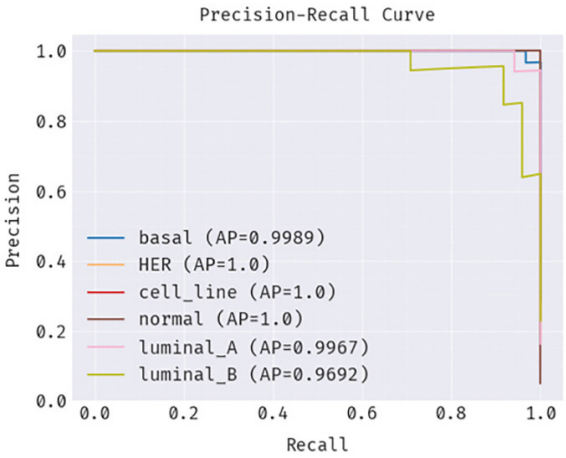

Fig. 7. Precision-recall curve for breast cancer training validation

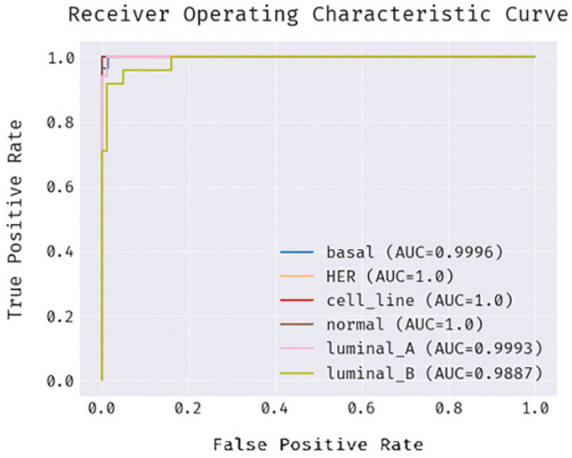

Fig. 8. ROC curve for breast cancer training validation

Table 2. Train and test validation for diabetes dataset

Parameter	Testing Values	Training Values
accuracy	0.9654	0.9572
Precision	0.9592	0.9578
Recall	0.9677	0.9464
Specificity	0.9677	0.9464
F1-score	0.9631	0.9517

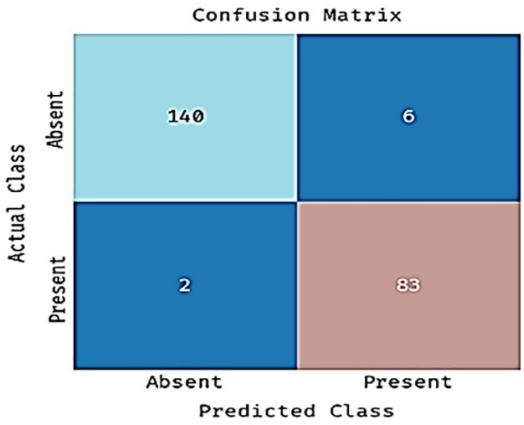

Fig. 9. Confusion matrix for diabetes testing validation

Figure 9 depicts a confusion matrix for diabetic testing validation, where the rows indicate the expected class and the columns reflect the actual class of data. The diagonally colored cells represent the tested systems that are categorized properly or inaccurately. The column just on the right-hand side represents each anticipated class, whereas the row at the bottom reflects each actual class's performance.

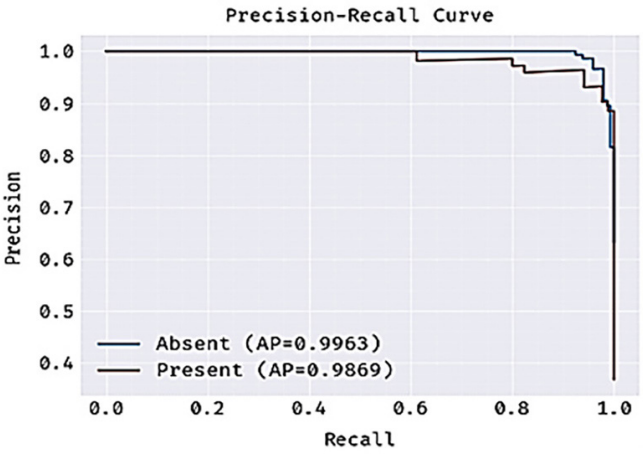

Fig. 10. Precision-recall curve for diabetes testing validation

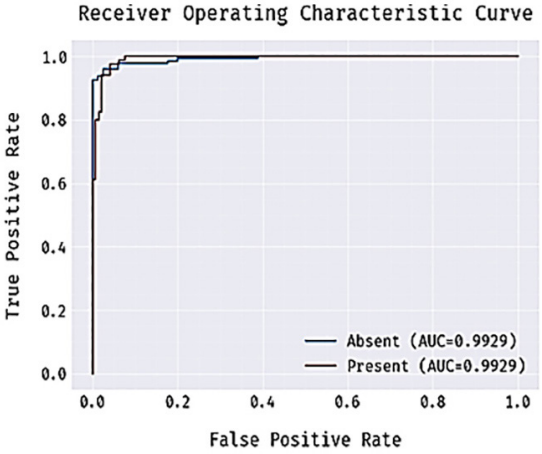

Fig. 11. ROC curve for diabetes testing validation

Figure 10 depicts the precision-recall curves for diabetic testing validation, where the x-axis represents recall and the y-axis represents accuracy. When the AP reaches 0.9963, it indicates presence, and when it reaches 0.9869, it suggests absence. Figure 11 depicts the ROC curve for diabetes testing validation, where the x-axis represents the false positive rate and the y-axis represents the true positive rate parameter. When AUC reaches 0.9929, it implies presence, and when AP reaches 0.9929, it shows presence.

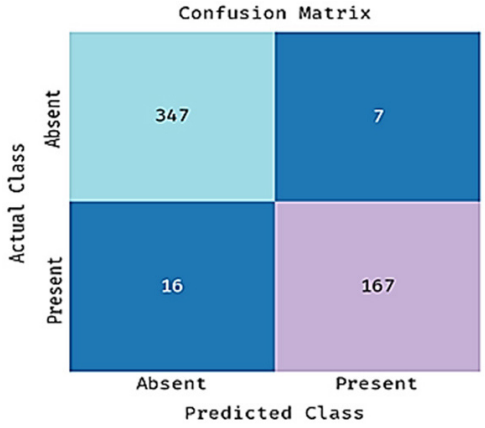

Fig. 12. Confusion matrix for diabetes training validation

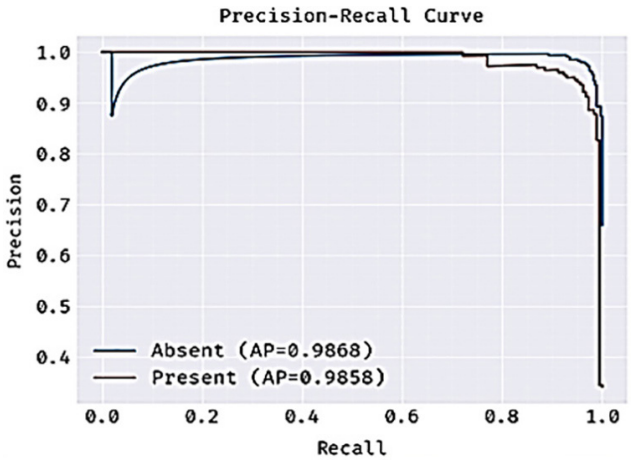

Fig. 13. Precision-recall curve for diabetes training validation

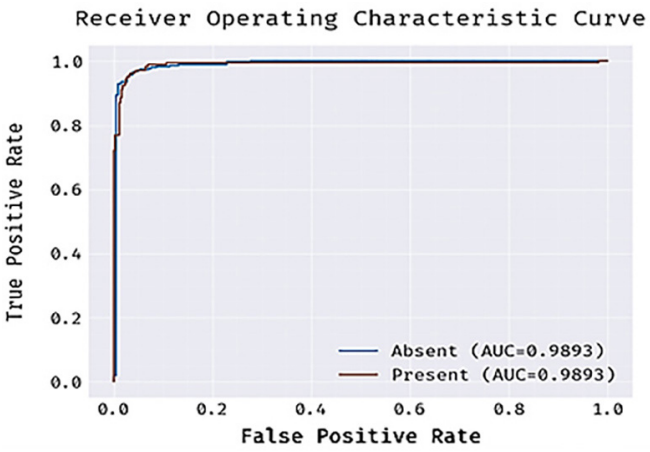

Fig. 14. ROC curve for diabetes training validation

Figure 12 depicts the confusion matrix for diabetic training validation, where the rows indicate the predicted class and the columns represent the actual class of data. The diagonally colored cells represent telecommunication that is properly and erroneously categorized. The column on the right-hand side represents each anticipated class, while the row at the bottom displays the performance of each actual class. Figure 13 depicts the accuracy-recall curve for diabetes training validation, where the x-axis represents recall and the y-axis represents precision. It is discovered that the AP is 0.9868 for both present and absence, indicating absence.

Figure 14 depicts the ROC curve for diabetes training validation, where the x-axis represents the false positive rate and the y-axis represents the true positive rate parameter. The AUC is found to be 0.9893 for both present and absent data. Table 3 shows training and test validation for the hepatitis dataset.

Table 3. Train and test validation for hepatitis dataset

Parameters	Testing Values	Training Values
accuracy	0.9605	0.9733
Precision	0.9912	0.9253
Recall	0.8	0.7984
Specificity	0.9417	0.9621
F1-score	0.8622	0.824

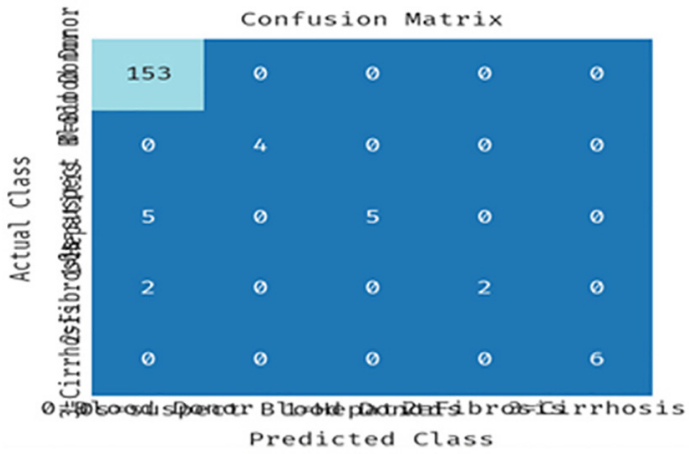

Fig. 15. Confusion matrix for hepatitis testing validation

The confusion matrix for hepatitis test validation is shown in Figure 15, where the rows represent the expected class and the columns represent the actual class of data. The diagonally colored cells represent the tested systems that are categorized properly or inaccurately. The columns on the right-hand side represent each anticipated class, while the row at the bottom reflects each actual class's performance.

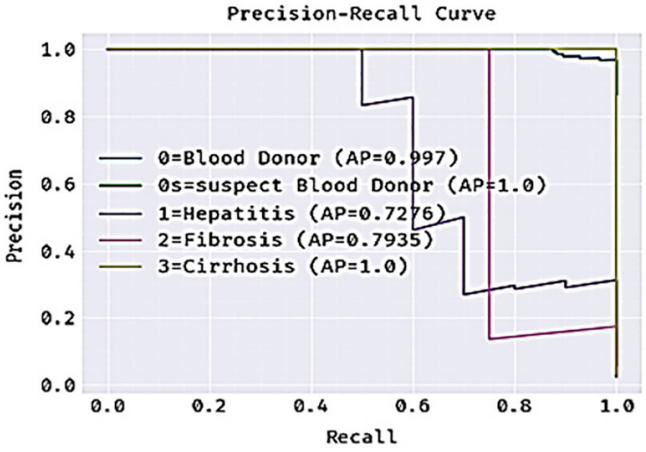

Fig. 16. precision-recall curve for hepatitis testing validation

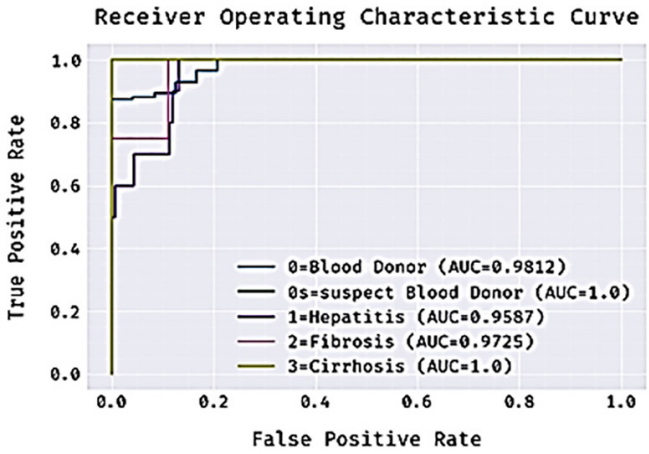

Fig. 17. ROC curve for hepatitis testing validation

Figure 16 depicts the precision-recall curves for hepatitis test validation, where the x-axis represents recall and the y-axis represents accuracy. AP is found to be 0.997 for blood donors, 1.0 for questionable blood donors, 0.7276 for hepatitis, 0.7935 for fibrosis, and 1.0 for cirrhosis.

Figure 17 depicts the ROC curve during hepatitis testing validation, where the x-axis represents the false positive rate and the y-axis represents the true positive rate parameters. The AUC is 0.9812 for blood donors, 1.0 for questionable blood donors, 0.9587 for hepatitis, 0.9725 for fibrosis, and 1.0 for cirrhosis.

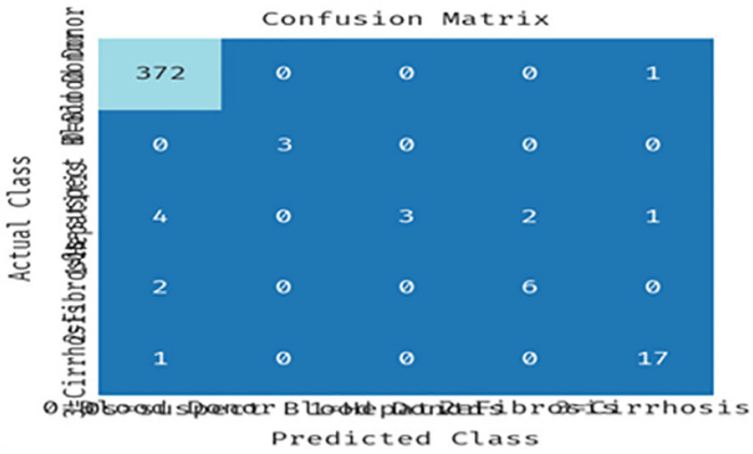

Fig. 18. Confusion matrix for hepatitis training validation

Figure 18 depicts the matrix for hepatitis retraining validation, where the rows indicate the predicted class and the columns represent the actual data class. The electric networks that are successfully and erroneously categorized are shown by diagonally colored cells. The columns on the right-hand side represent each anticipated class, while the row at the bottom reflects each actual class's performance.

Figure 19 depicts the precision-recall curves for hepatitis-trained validation, where the x-axis represents recall and the y-axis represents accuracy. The AP for blood donors is 0.9979, 1.0 for suspicious blood donors, 0.8145 for hepatitis, 0.8274 for fibrosis, and 0.979 for cirrhosis. The curve for hepatitis retraining validation is shown in Figure 20, where the x-axis represents the false positive rate and the y-axis

represents the true positive rate variable. The AUC is 0.9726 for blood donors, 1.0 for suspicious blood donors, 0.9005 for hepatitis, 0.9913 for fibrosis, and 0.999 for cirrhosis.

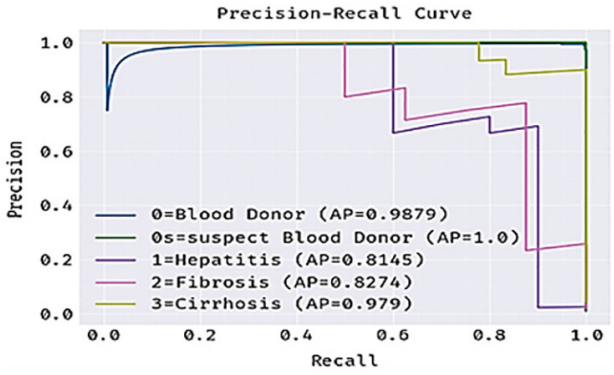

Fig. 19. Precision-recall curve for hepatitis training validation

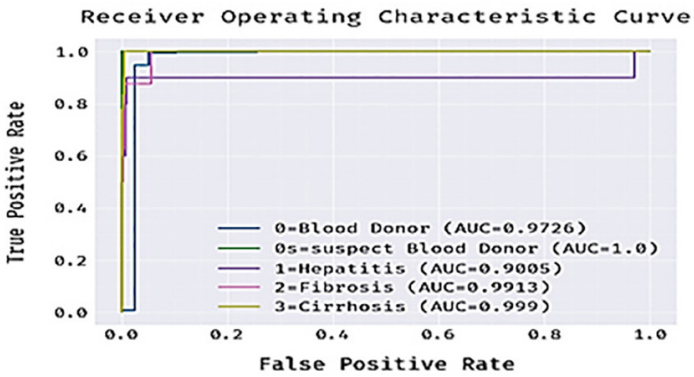

Fig. 20. ROC curve for hepatitis training validation

Table 4. Comparative analysis

Dataset	Parameter	Hybrid Approach	CNN Model	RNN Model
Breast Cancer	Accuracy	0.9783	0.9432	0.9367
	Precision	0.9861	0.9485	0.9402
	Recall	0.9833	0.9512	0.9426
	Specificity	0.9952	0.9356	0.9281
	F1-score	0.984	0.9501	0.9414
Diabetes	Accuracy	0.9654	0.9215	0.9124
	Precision	0.9592	0.9263	0.9185
	Recall	0.9677	0.9197	0.9113
	Specificity	0.9677	0.9176	0.9092
	F1-score	0.9631	0.9234	0.9151
Hepatitis	Accuracy	0.9605	0.9321	0.9246
	Precision	0.9912	0.9385	0.9302
	Recall	0.8	0.9256	0.9173
	Specificity	0.9417	0.9301	0.9224
	F1-score	0.8622	0.9324	0.9247

Table 4 provides a comparative analysis of the performance metrics for the hybrid approach, CNN model, and RNN model on three different datasets: breast cancer, diabetes, and hepatitis. For the breast cancer dataset, the hybrid approach achieves the highest values across all parameters, including accuracy (0.9783), precision (0.9861), recall (0.9833), specificity (0.9952), and F1-score (0.984). The CNN model and RNN model also perform well but consistently exhibit slightly lower values compared to the hybrid approach. Similarly, in the diabetes dataset, the hybrid approach outperforms the individual CNN and RNN models in terms of accuracy (0.9654), precision (0.9592), recall (0.9677), specificity (0.9677), and F1-score (0.9631).

The CNN model and the RNN model show slightly lower performance in all the metrics. In the case of the hepatitis dataset, the hybrid approach again demonstrates superior performance compared to the CNN and RNN models. The hybrid approach achieves higher values in accuracy (0.9605), precision (0.9912), recall (0.8), specificity (0.9417), and F1-score (0.8622). The results highlight that the hybrid approach tends to outperform the individual CNN and RNN models across all three datasets. It achieves higher accuracy, precision, recall, specificity, and F1-score values, indicating its effectiveness in handling these particular datasets. These findings suggest that combining the strengths of both CNN and RNN models in a hybrid approach leads to improved performance and better predictive capabilities for the given datasets.

5 CONCLUSIONS

This research focused on the categorization of microarray datasets using a hybrid classifier combined with optimization strategies. The suggested hybridization of CNN-HF gathers and classifies the microarray dataset for sources. Three conventional microarray cancer datasets, namely breast tumors, hepatitis, and diabetes, are utilized to validate the suggested technique. Null value handling and categorical-to-numerical techniques are employed as dimensionality reduction techniques to solve the curse of dimensionality and other challenges related to the nature of the data. The feature selection method is a hybridization of the mayfly and whale optimization algorithms. The binary cross-entropy is used to determine it because it is a conventional loss function and is recommended for classification problems. It has a large magnitude of error during both training and testing. We employed performance metrics such as classification results, accuracy, recollection, and classification error to evaluate the suggested method's effectiveness. We discovered that the proposed CNN HF achieves 97% accuracy, 98% precision, 98% recall, 99% specificity, and a 98.4% F1-score for the breast cancer dataset. It achieves 96% accuracy, 95% precision, 96% recall, 95% specificity, and a 96% F1-score for the diabetes dataset. It achieves 96% accuracy, 99% precision, 80% recall, 94% specificity, and an 86% F1-score for the hepatitis dataset. In the future, we want to improve the proposed method and apply it to multi-class microarray tumor databases. We also wish to improve the accuracy of classification on binary datasets that now have low accuracy.

6 ACKNOWLEDGMENT

This work was supported by the Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India.

7 REFERENCES

[1] Y. Guo, F.-L. Chung, G. Li, and L. Zhang, "Multi-Label Bioinformatics Data Classification With Ensemble Embedded Feature Selection," in *IEEE Access*, vol. 7, pp. 103863–103875, 2019. https://doi.org/10.1109/ACCESS.2019.2931035

[2] D. Chicco, "geneExpressionFromGEO: An R Package to Facilitate Data Reading from Gene Expression Omnibus (GEO)," In *Microarray Data Analysis. Methods in Molecular Biology*, Humana, New York, NY., vol. 2401, pp. 187–194, 2022. https://doi.org/10.1007/978-1-0716-1839-4_12

[3] C. Wang, Z. Gu, Y. Zhang, and J. Sha, "Screening of Short Single- and Double-stranded DNA Molecules Using Silicon Nitride Nanopores," *IEEE 17th International Conference on Nano/Micro Engineered and Molecular Systems (NEMS)*, Taoyuan, Taiwan, pp. 54–58, 2022. https://doi.org/10.1109/NEMS54180.2022.9791185

[4] Z.M.J. Sakhvidi, A. Zarei, V.R. Hachesu, and A. Zolfaghari, "Correction to: Evaluating the Relationship Between the Respiratory Exposure to the Benzene with the Primary Damages of Deoxyribonucleic Acid and Total Antioxidant Capacity in One of the Oil Companies in Iran," *Environ. Sci. Pollut. Res. Int.*, vol. 29, no. 32, p. 48347, 2022. https://doi.org/10.1007/s11356-022-19615-y

[5] A. Tyagi, S. Dubey, C. Sharma, P. Sudan, S. Rai, B.T.N. Kumar, M. Chandra, and M.A.K. Arora, "Complete Genome Sequencing and Characterization of Single-Stranded DNA Vibrio Parahaemolyticus Phage from Inland Saline Aquaculture Environment," *Virus Genes*, vol. 58, no. 5, pp. 483–487, 2022. https://doi.org/10.1007/s11262-022-01913-9

[6] U. Saha, Y.G. Srinivasulu, D.M.R.R.G. Soares, N. Madaboosi, and V.V.R. Sai, "Plasmonic Fiber Optic Absorbance Biosensor for MDR-Mtb Detection Using Padlock Probing," *Workshop on Recent Advances in Photonics (WRAP)*, Mumbai, India, pp. 1–2, 2022. https://doi.org/10.1109/WRAP54064.2022.9758216

[7] V.B. Canedo, N.S. Maroño, and A.A. Betanzos, "Feature Selection for High-Dimensional Data," *Progress in Artificial Intelligence*, vol. 5, pp. 65–75, 2019. https://doi.org/10.1007/978-3-319-21858-8

[8] P. Sova, Q. Feng, G. Geiss, T. Wood, R. Strauss, V. Rudolf, and N. Kiviat, "Discovery of Novel Methylation Biomarkers in Cervical Carcinoma by Global Demethylation and Microarray Analysis," *Cancer Epidemiology Biomarkers & Prevention*, vol. 15, pp. 114–123, 2020. https://doi.org/10.1158/1055-9965.EPI-05-0323

[9] M. Nosrati and M. Roushani, "Three-Dimensional Modeling of Streptomycin Binding Single-Stranded DNA for Aptamer-Based Biosensors, a Molecular Dynamics Simulation Approach," *J. Biomol Struct. Dyn.*, vol. 17, pp. 1–10, 2022. https://doi.org/10.1080/07391102.2022.2050945

[10] M.D. Catalina, K.A. Owen, A.C. Labonte, A.C. Grammer, and P.E. Lipsky, "The Pathogenesis of Systemic Lupus Erythematosus: Harnessing Big Data to Understand the Molecular Basis of Lupus," *J. Autoimmun*, vol. 110, p. 102359, 2020. https://doi.org/10.1016/j.jaut.2019.102359

[11] G. Shial, S. Sahoo, and S. Panigrahi, "Identification and Analysis of Breast Cancer Disease using Swarm and Evolutionary Algorithm," *IEEE Region 10 Symposium (TENSYMP)*, Mumbai, India, pp. 1–6, 2022. https://doi.org/10.1109/TENSYMP54529.2022.9864514

[12] L.R. Rodrigues, D.B.P. Coelho, and J.P.P. Gomes, "A Hybrid TLBO-Particle Filter Algorithm Applied to Remaining Useful Life Prediction in the Presence of Multiple Degradation Factors," *IEEE Congress on Evolutionary Computation (CEC)*, Glasgow, UK, pp. 1–8, 2020. https://doi.org/10.1109/CEC48606.2020.9185898

[13] G.T. Devi, V.T. Sujithra, P. Rajesh, and Francis, "Cancer MiRNA biomarker classification based on Improved Generative Adversarial Network Optimized with Mayfly Optimization Algorithm," *Biomedical Signal Processing and Control*, vol. 75, p. 103545, 2022. https://doi.org/10.1016/j.bspc.2022.103545

[14] R. Ye and P. N. Suganthan, "Empirical Comparison of Bagging-Based Ensemble Classifiers," *15th International Conference on Information Fusion*, Singapore, pp. 917–924, 2012.

[15] P. E. Jebarani, N. Umadevi, H. Dang, and M. Pomplun, "A Novel Hybrid K-Means and GMM Machine Learning Model for Breast Cancer Detection," *IEEE Access*, vol. 9, pp. 146153–146162, 2021. https://doi.org/10.1109/ACCESS.2021.3123425

[16] S. Das and P. N. Suganthan, "Differential Evolution: A Survey of the State-of-the-Art," in *IEEE Transactions on Evolutionary Computation*, vol. 15, no. 1, pp. 4–31, 2011. https://doi.org/10.1109/TEVC.2010.2059031

8 AUTHORS

Peddarapu Ramakrishna is a Research Scholar in Computer Science and Engineering department at the Koneru Lakshmaiah Education Foundation (KL Deemed to be University), Vaddeswaram, Andhra Pradesh, India. His research interests include Machine Learning, Artificial Intelligence, and Deep Learning. Currently, he is doing research in Biotechnology. He is a member of the CSI and ISTE is working as Asst. Professor in the Dept. of Computer Science and Engineering at VNR Vignana Jyothi Institute of Engineering and Technology, Bachupally, Hyderabad, Telangana. He is having 18 years of teaching experience and has published more than 15 research articles in ESCI and Scopus/international journals.

Dr. Pothuraju Rajarajeswari completed her Ph.D in Computer Science and Engineering and is currently working as Professor in KL University, Guntur, Andhra Pradesh. She has 22 years of academic teaching experience and has guided many M. Tech and PhD students. Her area of interest includes Artificial Intelligence, Intelligent systems, Data mining, Machine Learning, and Bioinformatics. She has published more than 65 research articles in SCI/ESCI and Scopus/international journals.