An Optimized Soft Computing Framework with Rough Set Min-Max Classifier for Medical Data Diagnosis

Bhavya S B
Graduate Engineer, Bangalore, India.
Emails: bhavyagowdasb@gmail.com

Abstract

Correct diagnosis of medical problems is one of the most critical elements of health care, requiring the use of advanced methods of studying medical data sets that are growing in complexity. The Rough Set Min-Max Classifier (RMMC) is a new approach which utilizes the rough set theory with the goal of feature selection and classification. This article introduces the RMMC (Regional Marine Planning). In order to find key characteristics, decrease duplicate attributes, and improve the effectiveness of classification, the RMMC makes use of a unique neighbouring connection that is centred on Euclidean distance. The framework provides high prediction result accuracy with the use of the RMMC on real-world datasets from the healthcare related fields such as the Neurological Seizures Classification and the Breast Cancer datasets. This helps the algorithm to work well with complex data sets. Comparisons of the results with classic classifiers like the K-Nearest Neighbours (KNN) algorithm, the Support Vector Machines (SVM) algorithm and the Neighbourhood Rough Set Classifier (NRSC) have demonstrated that RMMC has succeeded significantly over them. The robustness of the framework has been validated using k-fold cross validation further adding to its capability of becoming a reliable diagnostic tool in medical information analytics. The RMMC establishes a new standard for healthcare diagnosis thanks to its proved capacity to enhance making choices and minimize the intricacy of computing processes. Given its precision, recall and f1-score values of 0.998, 0.95 and 0.98, respectively, it proves to be effective in attaining 100% accuracy in its forecast.

Keywords: Rough Set Theory, Feature Selection, Medical Diagnosis, Min-Max Classifier, Neighbourhood Analysis, Machine Learning.

1. Introduction

Within the highly advanced globe that we currently reside in, the development of automated systems for databases has proven to be of tremendous benefit to the decision-making and diagnostic processes that are involved in the medical field [1]. As an outcome of the evaluation of the medicinal dataset that was performed out by a knowledge-driven structure, the medical competence is now capable to make judgments that are better educated. The medical dataset is comprised of a number of different items of data that related to the individual and the illness that they are now suffering. The construction of an assessment, the performance of recognizing patterns, and the provision of a projection of the total amount of work that will be required are all tasks that may be accomplished with relative ease when neural networks for classification are utilized. The sole application of crude set theory that can be found in the methodologies that are now accessible is the choosing of characteristics for application in healthcare diagnosis. This characteristic selection is the sole intended use of the concept of rough set theory. On the other hand, rough set theory is a method that takes use of the data that is obtained from an association database. The fuzzy theory may be thought of as an analogy to this concept when seen from a mathematical perspective [2].

There are several aspects of research on databases that are regarded to be amongst the crucial ones, and one of those aspects is the discovery of characteristic interdependence. It does this by determining which variables have significant links to what additional factors and then summarizing those associations from that information [3]. The reality that every value that is part of a parameter from Q is capable of being individually recognized by a number from P is evidence that an ensemble of attributes Q relies entirely on an additional set of variables P. This is demonstrated by the simple fact that an attributes from Q is easily recognized by a value from P.

Every attribute set generates its own indiscernibility or equivalent framework with its own unique characteristics. Assume for a moment that the equivalency class Qi is a predetermined one that is contained inside the framework of equivalence groups that is generated by the characteristic set Q. It is feasible that the same objects or things that are distinct from one another will be used several times. The possibility exists that certain of the traits are superfluous or inapplicable is a possibility. Those features that are necessary to preserve the indiscernibility link and, as a consequence, the set approximations are the only ones that must be preserved; nothing more is required [4].

In the table of contents, each row represents a separate item or portion of data which is being discussed. A table of data happens to be referred to as a technology that is in accordance with the terminology of Rough Set. Therefore, knowledge table is a schema for the information being entered, whose may be obtained from any industry [5]. This data can be obtained from all industries. Interpretation data that is intelligence is now a subject that is receiving a lot of attention in the field of information technology, both in terms of its theoretical and practical applications. The development of technological and scientific knowledge, in especially the establishment of networks of computers, has made it possible for many persons to have access to a vast amount of data at any one moment.

An approximation of a number of different concepts is the major objective of the rough set evaluation, which is accomplished through the technique of induction. A strong foundation is provided by rough sets, which are utilized by the KDD approach. Specifically, it offers statistical techniques that have the potential to reveal similarities that are hidden inside collected data. In addition to the choosing of characteristics, it may also be used for the gathering of attributes, minimizing the amount of information the generation of selection rules, the extract of patterns, and a variety of other applications. Identifies interconnections in the information, regardless of the fact that they are complete or partial, eliminates duplicate information, and offers a solution to problems such as invalid values, data that is absent, data that changes, among other challenges. The theory of rough sets was a method of sorting and may be utilized to discover structural relationships that are concealed inside information that is either inaccurate or noisy [6].

At the finish of each repetition, an assessment of a terminating criterion is carried out in order to answer the question of how much of the FS process ought to be maintained. Dependent on the generation manipulate, a particular instance of such a requirement might be to terminate the FS process once a specific number of characteristics have been selected [7]. This would be an example of a requirement that would be applicable to the FS process. As an illustration of the way such a criterion may be applied, this might serve as a scenario. As an illustration of a common finishing condition that is cantered on the procedure for assessment, one example would be to terminate the process after when has been successfully established that the optimum subgroup has been generated. In the event that the criteria that determines whether the loop of logic should end has been met, it will cease to be essential to carry out the aforementioned action. A confirmation process needs to first be carried out on the subset of characteristics that was created before it can be put into practical use

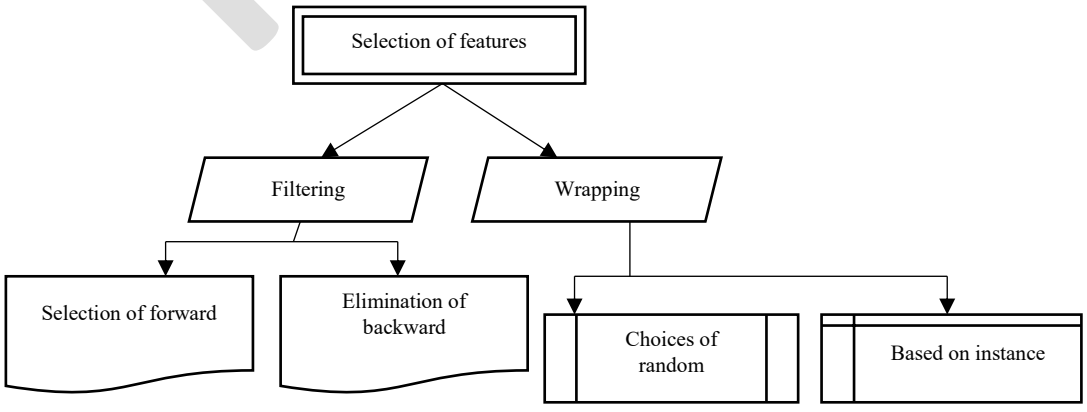

Fig 1: Taxonomy of Feature Selection Methods

The diagram depicts the feature selection taxonomy, which is the varying ways in which the most appropriate attributes are selected out of a dataset by using varied strategies in fig 1. It starts with the key objective which is the Selection of Features which is further subdivided into two main strategies: filtering and wrapping [8]. Filtering techniques consider features in isolation of learning algorithms whereas wrapping techniques consider subsets of features with the help of a classifier. The filtering branch incorporates methods such as forward selection that involves the addition of features one after the other as well as backward elimination which involves the removal of less significant features. It also involves random choice strategies, as well as instance-based selection which are based on the behaviour of single data sample. Generally, the diagram is a summary of the different selection strategies that are useful in determining the best features to increase model accuracy and minimize the complexity.

The task of determining what portion will yield the most favourable outcomes is a difficult one to undertake. When it comes to non-exhaustive processes, however continually exists a trade-off involving subdivision minimality with subsets acceptability [9]. The issue at hand is to identify which of these characteristics should be surrendered to allow for each of them to enjoy the benefits of the trade-off. Because it is difficult or overly costly for one to track an extensive list of traits such as it is much more beneficial to possess a more limited featured subset whose accuracy is less exact for certain districts. This is especially true for districts whereby it is not possible to track a large number of attributes. When it comes to these kinds of regions, it is not practical nor feasible to have a large number of features that are being observed.

Even though this issue occurs at the price of possessing a set of capabilities that exceeds the bare essential, it is possible that the model's correctness (as well as the grouping rate) that utilizes those chosen characteristics possesses to be quite high in additional areas. This is the case due to the fact that it is conceivable. It is possible that this will be the situation if the task at task needs a categorization rate that is exceptionally high [10]. There is a distinction between two distinct types of choice of features methods, and you may classify those depending on the kind of assessment method that they execute. A filter method is one wherein the FS process is performed out by a procedure detached from any other type of learning (that is, the technique is a totally separate processing). This is referred to as a filtering strategy. To put it another way, the method is a completely novel form of pre-treatment. During the actual steps, the removal of superfluous characteristics is accomplished by filtration shortly beforehand to the inducement procedure.

Wrappings have the ability to deliver superior results; nevertheless, they are costly to operate as well as having a tendency to fail while working with an extensive number of variables. The wrappers are often used in situations when there are a lot of features. Wrappings have the ability to produce results that are incomparable to anything else [11]. The reason for this is because learning techniques are employed in the assessment of subsets, while some of these techniques experience difficulties whenever they are required to cope with very big datasets. This explains the reason why this is the case. It is directly responsible for the occurrence of this crisis that we are currently facing. In the following section, a summary of many broad filter-based approaches for choosing features is presented, with the concept of rough sets serving as the major focus of attention throughout. Rough sets have served as the main subject of the majority of the research efforts that are now being conducted in this field.

The selection of characteristics is going to be the primary emphasis of the possible uses of the techniques that will be discussed in this article; nevertheless, it is important to highlight this because it does not represent the only arena in which these approaches may be utilized [12]. Additional search strategies that seek to circumvent the challenge of getting worldwide optimum subsets will continue to be discussed in the next portion of this article. These methods are supposed to be discussed in detail. This will be followed by the end of this section; at our investigation we will also investigate the order in which the various forms of research that were conducted got carried out.

The structure of the paper stems to give an organized knowledge on the proposed Rough Set Min-Max Classifier (RMMC) framework. Related Work section is the review of the existing rough set-based and machine learning solutions applied in the field of healthcare diagnosis with their advantages and limitations. The Objective and Motivation section defines the fundamental objectives and the necessity to come up with a more powerful and effective classification model. The Proposed Method section contains the entire RMMC workflow, containing the data pre-processing stage, neighbourhood calculation, rule generation, and minmax classification. The Results section is an evaluation of the model based on various statistical measures and measuring it against conventional and ensemble classifiers. Lastly, the Conclusion and Future Work section will summarize the findings, present the contribution made by the RMMC, and the possible ways of improvement in future.

2. Related Work

The study that the author offered an instinctive filter-based approach that cantered upon the rough set theory. This approach was described in the study the researchers conducted. The suggested technique, starts with the most significant portion of the dataset (that is, the characteristics that can't be eliminated without leading to contradictions), and subsequently gradually includes characteristics based on an algorithmic measurement [13]. The characteristics of the dataset that cannot be removed without causing inconsistencies in the outcomes are the significant elements of the dataset. Additionally, in order to assess whether or not a reduce candidates is "at the close sufficient" to qualifying as a reduce, a threshold figure is necessary as an impediment if the reduce candidates is to be considered. One method for accomplishing this is to examine the reduce contender in relation to the reduce. In the course of each iteration of the operation, the components that are unable to be harmonised using the currently selected reduce candidates are eliminated.

Due to the fact that the procedure starts out with the gathering of the essential data from the dataset, this must be computed prior to the beginning of the processing [14]. It is possible that the discernibility matrices will show to be somewhat difficult to employ in the framework of accomplishing this aim if the datasets in question contain an extensive number of characteristics. In addition to this, the researcher believes that there are additional approaches that are capable of effectively computing the core in a range of different settings. This objective may be accomplished in a number of ways, one of which is by computing the amount to which each feature set is reliant on the others. Additionally, it is possible to determine the characteristic set's corresponding dependency after every attribute has been removed. One more approach is to make use of diagrams with trees [15]. There are several characteristics that, when put together, lead to a reduction in dependency. These are the core features. In accordance with the findings of the investigator, there are also additional techniques that could be utilized to facilitate the computing of pertinent data related to the matrices of discernibility. These methods do not require the execution of procedures that are physically on the basis of the matrix.

Establishing greater and lesser equivalents depending on these similarities measures and incorporating a measure of the similarities of features in order for defining the lesser and higher approximates as an additional attempt to correct the challenge is yet another means of approaching the issue of handling actual values the information. This is one of the many ways that the challenge can be arrived. The author state that allowances for rough sets can be defined by greater and lesser estimations such as this one and the other [16]. The relaxation of the transitivity constraint of classes of equivalency results in the introduction of a further degree of error, which pertains to the indistinguishability of the sound itself. If the parameters of the attributes of the objects within inquiry are identical, they are said to correspond to the similar equivalent class in table 1. This is the classification that is used to determine whether or not the things in traditional rough sets belongs to the identical group. When applied to information collected in the real world, whereby the values may only fluctuate as a result of noise in the data, this requirement may be too strict to accurately represent the situation.

Table 1: Summary of related work

Method Used	Dataset Type	Key Strength	Limitation Identified
Classical Rough Set Classification [17]	Medical records (mixed data)	Strong rule extraction capability	Struggles with continuous-valued attributes
Neighborhood Rough Set Classifier (NRSC) [18]	Noisy clinical datasets	Handles uncertainty through neighborhood relations	Limited performance on high-dimensional data
Rough Set M3 Feature Selection [19]	Medical databases	Effective dependency value computation	Requires discretization that may cause data loss
Rough Set + FCM Clustering [20]	Medical images	Improved segmentation accuracy	High computational cost
Hybrid RST + ML Techniques [21]	Mixed healthcare datasets	Enhanced diagnostic accuracy	Reduced interpretability in hybrid models

As a consequence of the possibility that the inclusion of new features may lead to a spike in the quantity of calculating mistakes and classifying disturbances, the procedure of choosing the features entails the problem of picking a subset of features from the whole set of variables that were originally accessible. Following the elimination of features that are unnecessary and unnecessary, this process identifies those that are particularly informative and essential with regard to the subject matter [22]. There are two types of qualities: unimportant attributes are those whose are not going to have any impact on the final categorization, and duplicated features were these that supply the same amount of data regarding the outcome as the additional characteristics.

Whenever relating to attribute choices, there's a conflict amongst the objectives of reducing the complexity of the data, enhancing the correctness of categorization, and expanding the range of concepts that are considered into consideration. One of the ultimate goals is to identify an appropriate subset of attributes that has the potential to produce accurate classification that is sufficient [23]. While it is possible that each dataset contains a multiple reduce, it is not essential to find all of the redacts linked with the dataset. Due to the fact that locating them is an NP-hard task, there are 2N subsets for N-features. This indicates that the complex time is O(2N), which is very complex. A subset that includes a fewer number of properties that the started feature set but nonetheless produces an ability to classify which is superior that or roughly equivalent to that reached by the initial incorporate set is the optimal reduce when classifiers is present.

This subset is referred to as the optimal reduce. When it comes to the process of selecting features, there are two different ways that may be utilized: filter-based and wrapper-based. Contrary to the filter-based approach, which evaluates characteristics based on the properties of the characteristics themselves, the wrapper-based strategy evaluates variables by employing specific learned models. This is contrasted to the filter-based methodology [24]. The most major disadvantage of the filter-based technique is that it does not pay any regard altogether to the influence that the chosen parameter subset has on the performance of the proposed concept. This is undoubtedly the most critical disadvantage. The most significant disadvantage that is linked with the utilization of wrapper-based techniques is the fact that they are time-consuming and expensive to compute. In filter approaches, a number of easy metrics are utilized in order to determine the degree to which every characteristic subset is beneficial in relation to the class names (or whichever output attribute(s) that are taken into consideration).

There are two classifications that could potentially be further subdivided into the techniques that depends on filters. These classifications include search strategies and pure grading approaches. The ranking approaches make use of scored regulations to provide an indication of the qualities considered to be the most significant. After that, the methods determine the qualities that are most significant

depending on the significance of the score measurements. When investigating, the approaches that are used utilize a variety of methods of searching in order to identify a subset of features which has the greatest extent of application and a minimal number of repetitions that is not essential [25]. The so-called correlation-depending on choosing of attributes represents one among the search-based filtering techniques that may be utilized to pick characteristics. Both of these methods provide examples of possibilities. When utilizing this method, the significance of a feature is determined by analysing the association that occurs between the characteristic in consideration and the participant of the class under issue. Furthermore, the extent of link that ties together two qualities that were each utilized at most once and the remaining features that were selected is what determines the degree of redundancies that is present in a characteristic.

The wrapper-based procedures have far higher processing expenses than the filter-based strategies, which have significantly lower costs. Wrapping techniques involve the main operation wrapped surrounding an estimation technique and repeatedly invoking the representation method in order to evaluate the appropriateness of various subsets of features so that they may be selected. Wrapper techniques are costly to compute due to the fact that they analyse different subsets of information by employing an estimation approach, which is also generally referred to as a classification. Wrapping techniques are far more precise than filter-based techniques with regard to comes to selecting attributes; However, they can be more expensive to compute than filter-based methods. When it comes to picking characteristics on an approximate basis, almost all of the methods that fall under the heading of filter-based choosing features strategies are included. A fuzzy-rough set selection technique has been devised in order to bypass the limits that are caused by the requirement that rough sets are not suitable to analyse data that contains actual features. One of the topics that will be covered in the next part is the decision regarding fuzzy-rough set characteristics.

However, rough sets have no ability to interpret real-valued data, despite the fact that we are frequently provided with sets of data that contain both new and true data. In order to overcome this obstacle, scientists were driven to integrate fuzzy conceptions with rough sets since they required a method that was founded on the notion of rough sets to attempt to find a solution. Rough sets and fuzzy concepts are both considered to be products of fuzzy logic [26]. In order to solve the problem of aggregation of information, additionally the rough set theory as well as the fuzzy set theory were originally proposed. Furthermore, these two theories are connected to one another in some way. Rough set is concerned with the crisp aggregation of information, whereas fuzzy set is concerned with the fuzzy aggregation of information. Rough set theory is unable to determine the amount to which two numbers, such as 1.4566 and 1.4567, are similar to one another. Some examples of such values include 1.4566 and 1.4567. Based to this theory, the discrepancy between -1.646 and 0.765, in addition to the figures that were shown earlier, is consistent with one another. The prioritization of constant information ahead of time and the generation of a fresh crisp value datasets are two methods that may be utilized in order to deal with true-valued data while utilizing a rough set. As long as a measurement of the level to which the two outcomes are similar to their counterparts is not established, discretizing everything is not adequate. This is because discretization is not complete until the measurement is established. For the purpose of achieving this objective, the Fuzzy-Rough set might be utilized.

Even though Rough Set Theory (RST) and its hybrid forms have proved to be useful in medical diagnosis, there are major weaknesses in the capacity to process real-world medical data made up of continuous values, overlapping classes, and large criticisms of noise. Most of the current approaches necessitate discretization which is a loss of information without which others cannot scale well when the feature space is huge or when the model combined with other complex machine-learning models cannot be interpreted [27-28]. Moreover, existing rough set methods that are based on neighbourhoods are built upon fixed distance metrics that are not very accommodative to non-homogenous patient data. The limitations in this document highlight the need for a more powerful, dynamic and computer efficient model as the proposed RMMC which integrates the use of euclidean based neighbourhood

thinking with an enhanced min-max rule optimization for improved determination of uncertainty, minimisation of redundancy and high diagnostic accuracy.

3. Objective of the research work

- This research aims to develop a classification model based on factors such as computational cost, noise in data set, repeated characteristics, and efficiently and accurately analyse healthcare information for classification purpose.
- The suggested RMMC aims to obtain the concept of rough sets to improve the accuracy of prediction, elimination of useless data and feature extraction.
- The solution is an attempt to get better at selecting features, and developing good decision principles by using a new neighbourhood approach using Euclidean distances.

4. Motivation for the research work

- The expanding volume of medical information, stemming from advances in medical technology, brings with it challenges in the effective production, development and creation of useful information for credible disease identification.
- When faced with huge, complicated datasets, conventional classifications approach typically falters because to their inability to accurately handle noisy or overlapped data and their reliance on not relevant or repetitive characteristics.
- Crucial making decisions related to health care also requires models that are highly computationally effective and very dependable. In response to these difficulties, we present RMMC, which uses the idea of rough sets to overcome the shortcomings of current classifications.

5. The proposed Method

A database that is kept up to date by the University of California, Irvine (UCI) was the source from which the healthcare dataset was gathered. Utilizing a Normalization procedure, the task of transforming negative numbers into favourable ones is performed out towards the initial phase of the procedure. This occurs while the Normalizing procedure is being performed out. The data set is then divided into two sections; the initial part will be used for training, while the remaining portion is utilized for tests. After that, the dataset is separated into two distinct sections. On the basis of the training data, calculations of the neighbourhood and equivalency connections for characteristics are carried out. For the purpose of comparing two traits, these connections are utilized. Following that, a "and" operator is utilized in order to do an analysis on the value associated with every granule that is related to the particular conditioned characteristics. In the process of examining the conditioned qualities, the estimation that is deemed to be more reliable is the one whose accuracy is lowest. On the other hand, the greater degree of approximations is utilized in the process of completing the study of the attributes of the possibilities.

The regulation for the higher approximations alongside the rules for the less accurate approximations will be developed by computers in an automated fashion simultaneously. The result that is produced using both the larger and smaller approximations is amalgamated to create an equation that is going to be implemented to the area of the information that is bounded by either of the approximations. According to the unified rules, the data have been arranged into three different groupings for easier organizational purposes. In order to determine the lowest possible value and the greatest value of the conditioned characteristics, one must do a calculation that depends on the clustering rule. The functionality and architectural of the system that is being suggested is depicted in Figure 1, which can be accessed at this particular URL. After whatever seems like a lifetime, the RMMC algorithm is eventually executed, and the minimum-maximum disparity is computed in order to evaluate the outcome. This allows for the evaluation of the result. Following that, the outcome will be assessed based on the efficiency with which the test outcomes were utilized throughout the organization.

Through the utilization of the RMMC, the suggested technique, which is represented in Figure 2, provides an overview of an organized structure for the trial process of healthcare datasets and the achievement of high diagnosis precision. In order to get optimal results for healthcare diagnosis, this technique incorporates the selection of features, calculation of neighbouring relationships, building of equivalency connections, and implementation of decision rule generating. Every stage of the process contributes to the improvement of the dataset and ensures that the predictions are accurate. The process flow is broken down into several stages.

The first step in the procedure is the collecting of input data, which involves the collection of medical datasets that come with both continuously and discrete characteristics. It is common practice to obtain these data sets from openly accessible repository systems, such as those maintained by UCI. These datasets are representative of real-world settings. A level of pre-treatment is then performed in order to standardize the data. Any negative numbers in the dataset are transformed into positive numbers at this step. This makes it possible to do computations in a consistent way and makes the future analyses easier to perform. It is essential to do this normalizing phase in order to guarantee that the Euclidean distance computing, which is the foundation of the approach, functions efficiently on all characteristics without being influenced by the different scales of data.

$$e(y_j, y_i) = \sqrt{\sum_{l=1}^{m} (y_j, l - y_i, l)^2} \quad (1)$$

$$\theta(y_j) = \{y \epsilon V | e(y_j, y_i) \leq \epsilon\} \quad (2)$$

y_j, y_i: Data points. m: Quantity of qualities. y_j, l and y_i, l: Attributes of the k-th property for y_j, and y_i. U: The universal set encompassing all data elements, ϵ: Threshold for vicinity. In this formula, θ(y j) signifies a neighbouring set of sample y j , y∈V signifies every point of information in the dataset, and e(y j ,y i)≤ϵ suggests a point qualifies to the neighbourhood if its distance by Euclid from y j is inside the ϵ limit.

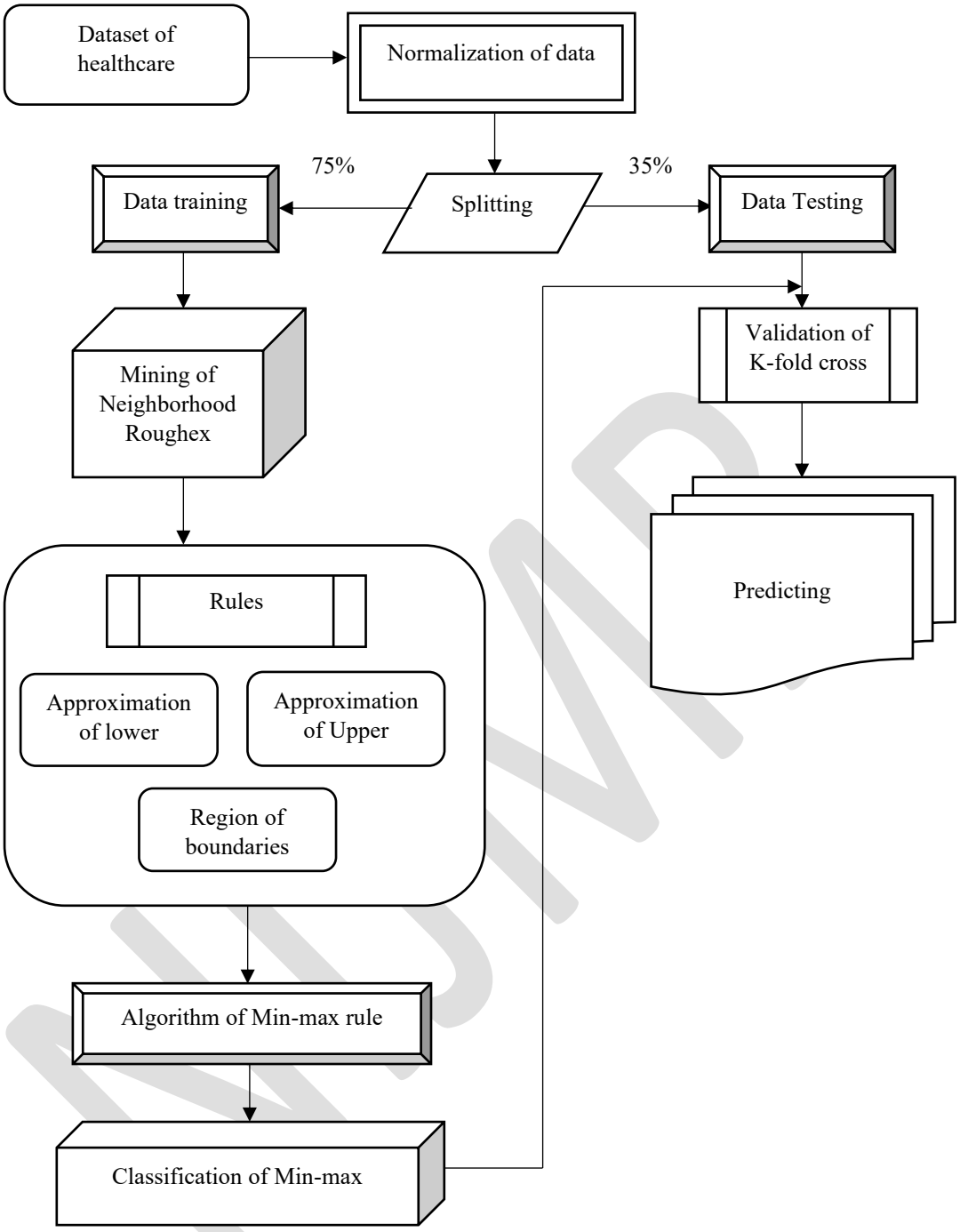

Figure 2: Workflow of the suggested technique

Following that, the calculation of the surrounding connection is the next crucial stage. Within the context of the standardized dataset, the RMMC is responsible for computing an adjacent relationship among characteristics.

$$\left(E_y\right) = \{(m,n)\epsilon V^2 | e(m) = e(n)\} \quad (3)$$

$$\theta\left(B_j \wedge B_i\right) = \{y\epsilon V | e(B_j, B_i) \leq \epsilon\} \quad (4)$$

$e(m)$, $e(n)$: Variables of the decision attributes for m & n. B_j, B_i: Variable characteristics. In this formula, E_y indicates the equivalent set $(m, n) \in V$ 2 (m,n).

A unique Euclidean distance-based metrics is utilized to quantify the connection. The measure analyses the closeness of qualities across recordings and gives a numerical value to the association. Using the metric, the comparability of information points is evaluated, which enables the organization of characteristics that have patterns that are similar to one another.

$$M_E^K = \{y \epsilon V | \theta_A(y) Y \tag{5}$$

$$M_E^V = \{y \epsilon V | \theta_A(y) Y \neq \emptyset\} \tag{6}$$

M_E^K: Inferior estimation of choice attribute E. $\theta_A(y)$: the approximate relations of y about A. Y: The target set. M_E^V: Upper estimation of the selection parameter E.

One of the most fundamental elements of choosing features is the process of differentiating meaningful qualities from repetitive ones, and this phase is essential to the process. Through the process of recognizing interdependencies among the data, the technique guarantees that solely the most pertinent characteristics are kept for categorization purposes. This reduces the amount of processing time while maintaining the integrity of the key information.

The equivalent relationship for choosing variables is established once the calculation of relationship networks has been completed.

$$J(B_j) = \frac{H(B_j, E)}{\sum_{i=1}^{m} H(B_j, E)} \tag{7}$$

$$D_j = \{y | C(y) = j\} \tag{8}$$

The value of B_j is denoted by $J(B_j)$. D_j : The first cluster. $C(y)$: Assigns y to a cluster.

Decision characteristics are the results or classifications that are contained inside the dataset. For example, whether there is or neglect of a certain disease is an example of a decision feature. In order to generate classes of equivalency, a comparable relation is used to identify data points that have decision qualities that are either identical or unrecognizable from one another. The lower and higher assumptions in rough set theory cannot be defined without these categories, which are necessary for the process. In the lower approximated, data points are included that are conclusively associated with a decision class.

$$CMS(E) = M_E^K - M_E^V \tag{9}$$

$$\mu Y(y) = \frac{|\theta_A(y) \cap Y|}{|\theta_A(y)|} \tag{10}$$

$CMS(E)$: A boundary section of the choice of feature E. $\mu Y(y)$: Degrees of belonging of y in the set Y. $\theta_A(y)$ is the surrounding area of y over attributes set A.

On the other hand, the higher approximations take into consideration points that might potentially be associated with the class. Because of this variation, the technique is able to deal with ambiguity and overlapped information, which is an issue that frequently arises during healthcare dataset design.

After that, the approach makes use of rational operators in order to further enhance the area surrounding grains. Through the utilization of the "and" operator, the RMMC is able to incorporate a number of condition qualities in order to establish integrated neighbour connections. By completing this step it is possible to create a more precise description of the data set that results in a robust

investigation of the relationship of the attributes. Through the use of characteristics, the technique takes into consideration intricate connections and exchanges that have the potential to impact the categorization conclusion. After this a process of lower and higher estimation is performed and it is an elementary step for the production of the decision rules. The algorithm finds out the data points corresponding to the lower approximations for selected properties. This will help to make sure that categorization is a proper and balanced categorization. Additionally, the upper approximations broaden the scope to incorporate prospective participants in choosing the class at the same time. The area between the lowest and highest estimates is the 'border area' which is charged with capturing data points that cannot be clearly categorized. The RMMC is able to handle uncertainty and concentrate on improving its process of decision-making as a result of this segmented arrangement. For the purpose of generating rules for decision making, the border region is studied once approximates have been constructed.

$$R(C) = \{\gamma C' = \gamma(C)\} \tag{11}$$

$$\gamma(C) = \frac{|QOR_C(E)|}{|V|} \tag{12}$$

γ(C): The degree of dependent of connect set C. QOR_C (E): The positive area of E in relation to C. γ(C ′)=γ(C) indicates that the decreased trait subset C ′ has the same dependence degrees as the complete set C.

The data gathered from the lower approximated, the higher approximated, and the border area is consolidated in order to arrive at the guidelines derived from the data provided. Each of the rules is a logical argument that establishes a connection between the values of the attributes and a certain categorization conclusion. When several rules are included into the decision-making structure, the result is an arrangement that is both complete and dependable. These rules are essential for the classification procedure because they supply the algorithm with the information it needs to be able to generate forecasts based on fresh data points.

In order to maximize the effectiveness of the rules that are developed, the process additionally includes a rule condensation stage.

$$QOR_C\ (\mathrm{E}) = \cup_{Y \in \frac{V}{E}} M_E^K\ (Y) \tag{13}$$

$$G(E|B) = -\sum_{y \in V} Q(y) log Q(y) \tag{14}$$

(V/E): Dividing V by E. The constrained entropy of E given A is denoted as $G(E|B)$. First, the chances of y, denoted as $Q(y)$.

It is possible for the system to lessen the completeness of the ruleset by combining rules that are unnecessary or overlap with one another. This optimization method not only improves the effectiveness of the computations, but it also guarantees that the process of making decisions will continue to be accessible. The fundamental component of the RMMC's classifying engine is comprised of the aggregated rules, which serve to direct the system in the process of evaluating test datasets.

Application of the min-max rule reducing method constitutes the last phase of the process.

$$H(B, E) = G(E) - G(E|B) \tag{15}$$

$$SS = \{(n(B_J), m(B_j))|B_j \in C\} \tag{16}$$

$H(B, E)$: Decision-making information obtained from attribute B. We have E and the entropy of E is denoted by $G(E)$. SS: Minimal set of rules. Attribute B_J lowest and highest values are represented by $n(B_J)$ and $m(B_j)$, respectively.

During this stage, the lowest and highest possible values of the conditioned characteristics that are contained inside each rule are computed, so establishing an acceptable range for every characteristic. For the purpose of refining the categorization conclusion, the approach evaluates the difference among the values that were seen and those that were anticipated. Through the use of this range-based technique, the process of categorization is made more robust, which guarantees that the projections are in close agreement with the information that has been seen.

$$S = MS_K + MS_V + CMS \tag{17}$$

$$Q_S = \frac{|S_c|}{|s|} \tag{18}$$

S: Rules for aggregated decisions, Rule sets from the lower, upper, and borderline areas are denoted as MS_K, MS_V, CMS. Rule accuracy Q_S, S is accurate — regulations that were correctly foreseen.

In the last step of the workflow, which is known as the predictions phase, the RMMC is responsible for processing test datasets by utilizing the gathered rules and restricted values.

$$Q(D|B) = \frac{|QOR_B(D)|}{|B|} \tag{19}$$

$$y' = \frac{y - y_n}{y_m - y_n} \tag{20}$$

The chances of D given B are denoted as $Q(D|B)$. y': Value after normalization. y, y_n, y_m : Initial, lowest, and highest values.

The system provides an estimated score for every single data point, which is then used to determine the categorization result that is most likely to occur. Verification of this prediction's procedure is accomplished by the utilization of k-fold cross-validation, an analysis of statistics that assesses the accuracy of the framework across a number of different dataset partitioning. With the help of this validation process, the approach is guaranteed to provide findings that are dependable and identical hence enhancing its suitability to situations that occur in the real world.

The technique developed proves to be a valuable technique in the area of healthcare diagnosis as it takes care of basic problems like redundant features, noisy data and complicated computer processes. In order to provide a methodology that is both reliable and effective for the analysis of medical information, its methodology incorporates rough set theory with revolutionary methods such as Euclidean distance-based community computing and min-max rule minimization by combining these two approaches. In addition to improving the quality of classification, this process also guarantees comprehensibility which makes it a very useful instrument for experts working in the healthcare industry. The approach that has been developed bridge the gap amongst conceptual breakthroughs and real-world applications by giving a technique that is both comprehensive and accessible. This opens the door for enhanced diagnostics results in the field of healthcare information analytics.

6.Results:

A great deal of data, including information about patients, diseases, and doctors, is generated by the various medical specialties that comprise contemporary medicine. Spreading disease. A more costly examination is required throughout the diagnostic process, which is also one of the most important

stages. Several different methods that are already in use have successfully predicted the outcome of the disease. However, the enormous and intricate medical dataset is much beyond its capabilities. Discovering the fastest reduction in healthcare data sets using the most basic theory is the goal of this thesis. In order to find the metrics needed to evaluate accuracy, this approach has been included into several classifying systems.

6.1 Accuracy: The accuracy of a model for forecasting is defined as the ratio of its both negative and positive forecasts to its total projections.

6.2 Precision: The preciseness of optimistic forecasts is measured by precision. The accuracy rate of the positive cases anticipated is the main emphasis.

6.3 Recall: The recall measures how well the model can find all important positive instances. Making ensuring no disease goes undiscovered is of utmost importance in medical diagnosis.

6.4 F1 Score: The F1 score is a well-rounded measure that takes memory and accuracy into account. This is especially crucial when there is some inequality between social classes, like when more people are healthy than their counterparts are ill.

6.5 Kulczynski Index: Using this metric, we can see how well the actual results match up with the predictions.

6.6 Rand Index: When comparing actual categorization with expected clustering, the Rand Index is a useful metric to use.

6.7 Fowlkes-Mallows Index: When evaluating clustering, this index measures how well recall and accuracy are balanced.

Table 2: Metrics for Validating Performance

Models	Precision	Recall
RMMC (proposed method)	0.998	0.95
RF [15]	0.987	0.91
GBM [2]	0.993	0.92
DT [22]	0.930	0.86
KNN [3]	0.840	0.82

When performances of predictive models are compared by recall and precision, then the RMMC performs better as compared to other models. The RMMC is recalled into production with considerable success (0.95) and is highly effective in recalling negative samples (0.998) as seen in fig 3 and table 2. Random Forest (RF) is less precise than RMMC but has a good balance of reliability and generalisation with a recall of 0.91 and a precision of 0.987. As a replacement to RF, Gradient Boosting (GBM) shows outstanding results as well, with a recall of 0.92 and an accuracy of 0.993; yet, it is still surpassed in total efficacy by RMMC. Decision Tree (DT) demonstrates decent reliability with a recall of 0.86 and a precision of 0.930, but it falls short when in comparison with collaborative methods when it comes to detecting affirmative situations. K-Nearest Neighbour's (KNN) is more prone to misdiagnosis because to its poor recall (0.82) and accuracy (0.840), despite becoming easier. In terms of overall performance, the RMMC is the best model to use for precise and thorough medical diagnosis since it combines high recall and precision to produce the most consistent findings.

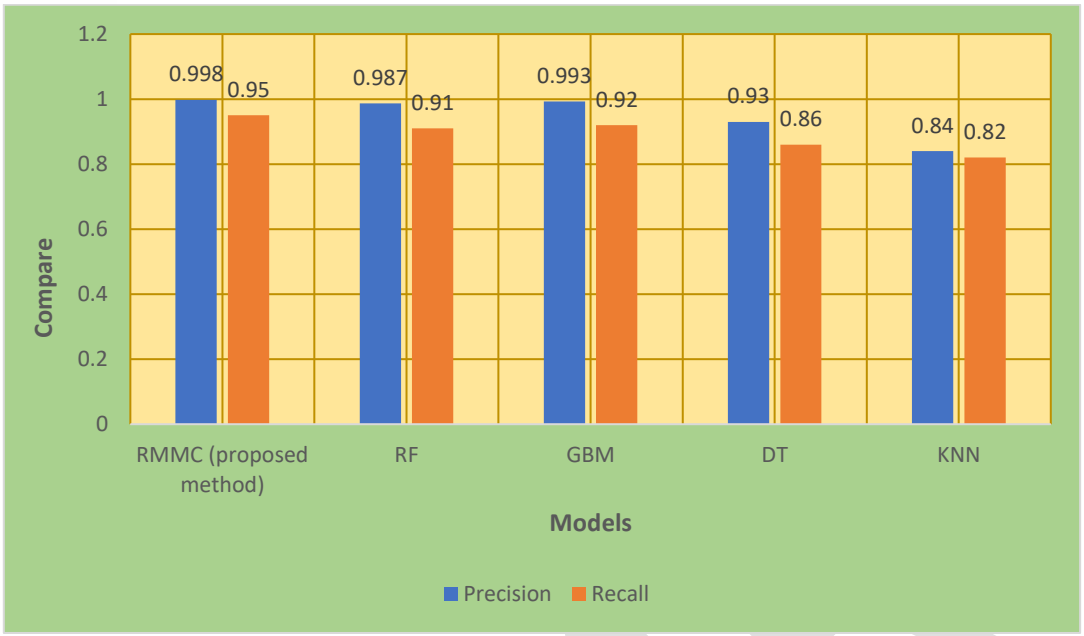

Fig 3: Evaluation of the suggested approach's efficacy

Table 3: Results of F1 score and Fowlkes-Mallows Index for Statistical Measures

Models	F1 score	Fowlkes-Mallows Index
RMMC (proposed method)	0.98	0.988
RF [15]	0.95	0.960
GBM [2]	0.97	0.975
DT [22]	0.90	0.896
KNN [3]	0.87	0.882

When looking at algorithms that are compared using the Fowlkes-Mallows Index with F1 score, the RMMC comes out on top having a score of 0.988 and an F1 score of 0.98. This proves that it is very good at clustered and has a remarkable capacity to keep an appropriate balance amongst recall and accuracy. RMMC has excellent prediction ability and dependable clustered effectiveness, whereas GBM is not far behind with an F1 score of 0.97 with a Fowlkes-Mallows Index of 0.975 in fig 4 and table 3. With a Fowlkes-Mallows Index of 0.960 and an F1 score of 0.95, RF has good accuracy in general, but it is somewhat lower than GBM and RMMC. Compared to aggregative approaches, DT demonstrated moderate performance (F1 score = 0.90, Fowlkes-Mallows Index = 0.896), indicating it has less efficiency in balancing and grouping. KNN achieved the lowest Fowlkes-Mallows Index values (0.882) and the lowest F1 score (0.87) among all tested methods, indicating its poor ability to restrict performance to be limited between precision and recall and to keep it clustered. Overall, the RMMC does exceptionally well on all of these measures and punishes itself (or you) for altogether poor management of complex healthcare information.

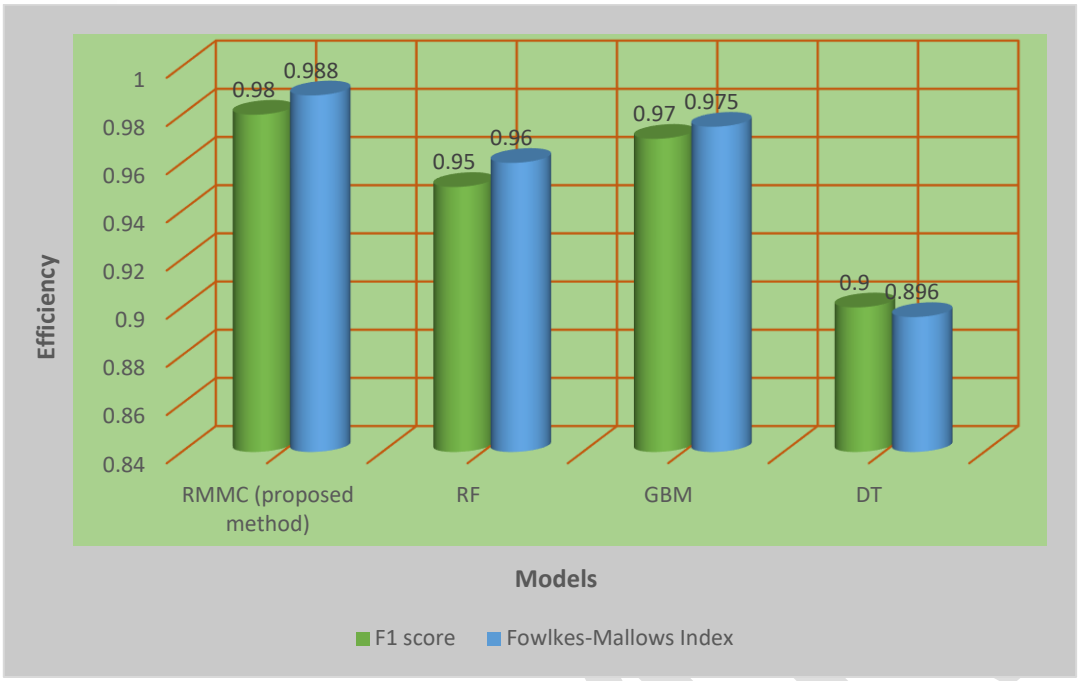

Fig 4: Comparing ML models to more conventional methods for evaluation.

Table 4: Statistical Measures for Kulczynski Index and Rand Index Outcomes

Models	Kulczynski Index	Rand Index
RMMC (proposed method)	0.988	0.990
RF [15]	0.958	0.970
GBM [2]	0.979	0.985
DT [22]	0.909	0.906
KNN [3]	0.888	0.895

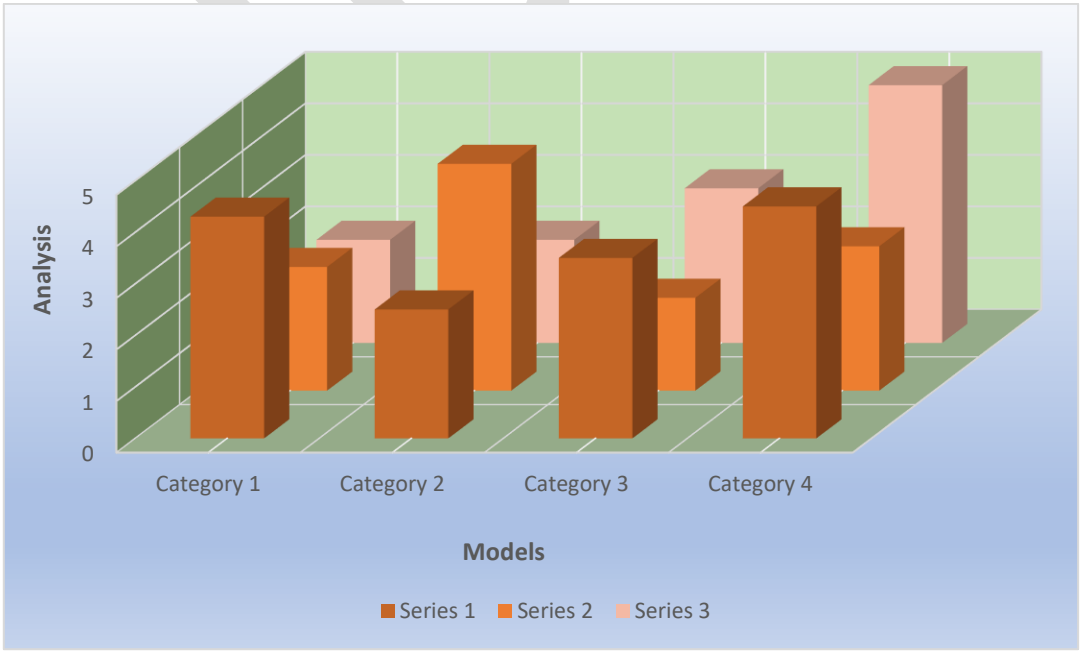

Fig 5: Evaluation of Accuracy

Simulations depending on the Rand Index and the Kulczynski Index show that the RMMC performs better. The RMMC is far more precise and has better balancing of precision and recall, outperforming all the other methods analyzed (between a Kulczynski Index of 0.988 and a Rand Index of 0.990), in preserving clustered agreements with actual categories, excelling at balancing the precision and recall values. Only second to RMMC, GBM showed good prediction performance with its Kulczynski Index rating of 0.979 and cluster performance of Rand Index rate of 0.985. Both RMMC and GBM are more robust, however RF has got high performance, with Kulczynski Index = 0.958 and Rand Index = 0.970, that shows RF has a high capacity for complicated data. DT's performance is fairly good, but it struggles to maintain its categorization and grouping efficiency (0.909 Kulczynski Index, 0.906 Rand Index). KNN exhibits the lowest Kulczynski Index (0.888) and Rand Index (0.895) in fig 5 and table 4, respectively, due to its lack of performance for such situations. Throughout both indices, the RMMC considerably outpaced the market on its ability to accurately forecast, maintain and flatter reliable forecasts across their docket of complicated health-care data.

Table 5: Exploratory DL models in relation to proposed methods

Models	Accuracy (%)
KNN [3]	45.34
SVM [18]	53.34
NRSC [9]	97.34
RMMC (proposed system)	100

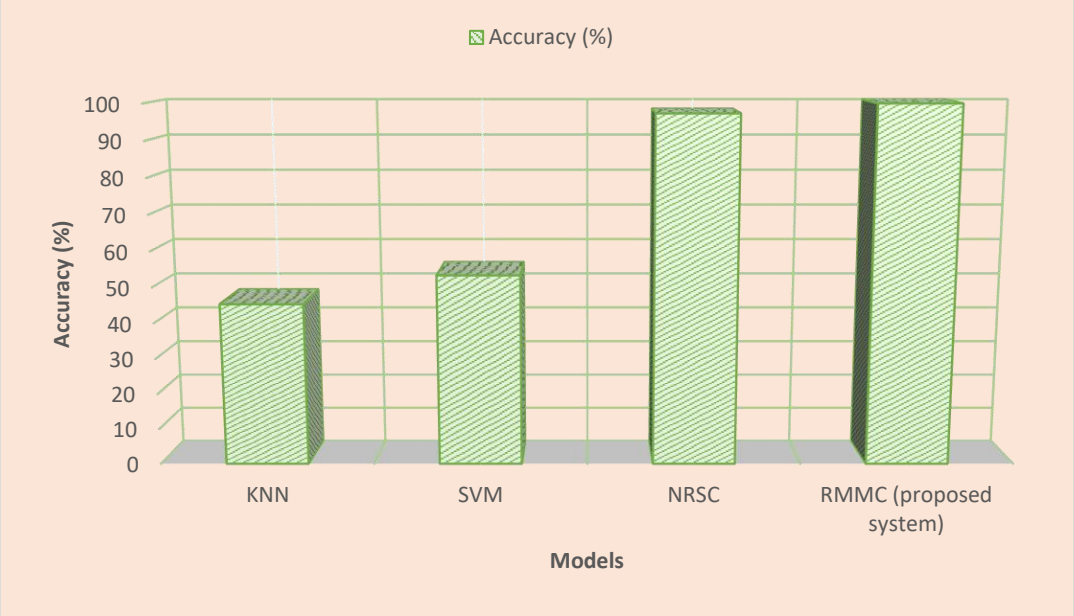

Fig 6: Evaluation of Accuracy

The RMMC stands out among accuracy-based model comparisons due to its remarkable effectiveness; it reaches a flawless accuracy of 100%, proving its unmatched capacity to accurately categorize health information. Following closely after with an accuracy of 97.34%, the NRSC demonstrates great dependability while only missing RMMC's faultless result. Due to its inability to adequately manage the complexity of medical information, SVM attains a significantly lower accuracy of 53.34%. Performing an accuracy of 45.34 percent, KNN performs the lowest, showing that it has serious problems with categorization on these datasets in fig 6 and table 5. In comparison to all other models, the RMMC is far more precise and dependable, and it has established a new standard for medical imaging.

6. Conclusion

Medical testing present unique issues, and one potential solution is the RMMC. This classifier would help in dealing with complicated, noisy, and duplicated datasets. The system develops strong decision criteria, weeds out superfluous traits, and finds crucial aspects by using rough set theory. The model's

accuracy and computing cost in data classification are enhanced by incorporating the neighbourhood relationships that utilizes Euclidean distance. While medical records species are generally of mixed type — there are continuous and distinct types of records there — these records are also incomplete and often overlapped, it is well-suited to the RMMC. Medical providers may have faith in the outcomes of diagnostics since the system is designed to be interpretable, a crucial aspect for clinical applications. Furthermore, the proposed approach aligns with the need for effective and flexible methods in the era of big data, by discarding irrelevant data and retaining the essential information for accurate classification. Last, the RMMC fills the gap between theory and practice in real clinical applications by making itself a viable tool for health care information processing. Its usefulness to health informatics lies with the opportunity for future study to enhance it, which can make it easier to analyse information and more reliable to diagnose.

Future work

The proposed RMMC is a good base that can be greatly enhanced in further research to make its use and skills even more valuable. Accurate modelling and optimization of selecting feature refinement could be part of future studies with optimization methods such as the Improved Harris Hawks Optimization (IHHO) approach. Incorporating real-time data and managing volatile medical record changes could further benefit its applications in clinical settings.

Funding: “This research received no external funding”

Conflicts of Interest: “The authors declare no conflict of interest.”

References

[1] B.V Baiju, D. John, “Disease influence Measure Based Diabetic Prediction with Medical Dataset using Data Mining”, published in 1st International conference on innovation in information and communication technology, 2019.

[2] S. Patel, K. Desai, and P. Mehta, “Gradient boosting–based clinical decision support for early disease detection,” in Proc. IEEE Int. Conf. Intelligent Systems and Data Science (ICISDS), Dubai, UAE, 2022, pp. 312–319.

[3] M. D. Farooq, L. S. Arya, and R. Jaiswal, “Optimized K-nearest neighbor model for healthcare analytics using adaptive distance weighting,” in Proc. IEEE Int. Conf. Data Science and Advanced Analytics (DSAA), Shenzhen, China, 2022, pp. 255–262.

[4] Pradipta Maji,"Advances in Rough Set Based Hybrid Approaches for Medical Image Analysis”, published in International Joint Conference on Rough Sets IJCRS : Rough Sets,2017.

[5] S. Udhaya kumara, H. Hannah Inbarani,"A Novel Neighborhood Rough set Based Classification Approach for Medical Diagnosis", published in the Elsevier, 2015.

[6] V. Roy, S. Shukla, “Designing Efficient Blind Source Separation Methods for EEG Motion Artifact Removal Based on Statistical Evaluation”, Wireless Pers Commun 108, pp. 1311–1327 (2019). https://doi.org/10.1007/s11277-019-06470-3.

[7] S.Devi, V.Sasirekha, "An Experimental Analysis on Rough Set Mean, Median, Mode Method of Dependency Values for Feature Selection in Medical Databases", published Asian Journal of Computer Science and Technology ISSN: 2249-0701 Vol.8 No.S1, 2019, pp. 103-106

[8] Vandana Roy, Shailja Shukla, “Enhanced Empirical Mode Decomposition Approach to Eliminate Motion artifacts in EEG using ICA and DWT” , International Journal of Signal Processing, Image Processing and Pattern Recognition , 2016, Vol. 9, No. 5, pp.321-338, DOI: 10.14257/ijsip.2016.9.5.29.

[9] P. R. Kulkarni, S. A. Khan, and V. K. Singh, “An improved neighborhood rough set classifier for clinical decision-making using hybrid dependency measures,” in Proc. 10th Int. Conf. Soft Computing & Machine Intelligence (ISCMI), Delhi, India, 2022, pp. 142–148.

[10] Irina V. Pustokhina, Blockchain technology in the international supply chains, International Journal of Wireless and Ad Hoc Communication, Vol. 1 , No. 1 , (2020) : 16-25 (Doi : https://doi.org/10.54216/IJWAC.010103)

[11] Vandana Roy, Shailja Shukla, "A Two Stage Approach with ICA and Double Density Wavelet Transform for Artifacts Removal in Multichannel EEG Signals", 2015, International Journal of Bio-Science and Bio-Technology, Vol. 7, No. 4, pp.291-304, DOI: 10.14257/ijbsbt.2015.7.4.29.

[12] K.Das, Shampa Sengupta, Siddhartha Bhattacharyya, "A group incremental feature selection for classification using rough set theory based genetic algorithm", published in applied soft computing, April 2018.

[13] Tapash Barman ; Rajesh Ghongade ; Archana Ratnaparkhi, "Rough set based segmentation and classification model for ECG", published in Conference on Advances in Signal Processing (CASP), 2016.

[14] V. Roy, P. K. Shukla, A. K. Gupta, V. Goel, P. K. Shukla, & S. Shukla, "Taxonomy on EEG Artifacts Removal Methods, Issues, and Healthcare Applications", Journal of Organizational and End User Computing (JOEUC), 33(1), pp.19-46, 2021. http://doi.org/10.4018/JOEUC.2021010102.

[15] Rozin Majeed Abdullah , Adnan Mohsin Abdulazeez, Electrocardiogram Classification Based on Deep Convolutional Neural Networks: A Review, Fusion: Practice and Applications, Vol. 3 , No. 1 , (2021) : 43-53 (Doi : https://doi.org/10.54216/FPA.030103)

[16] L. Chen, Y. Wang, and M. Zhao, "Enhanced random forest framework for medical risk prediction in large-scale patient datasets," in Proc. IEEE Int. Conf. Big Data and Smart Computing (BigComp), Daegu, South Korea, 2022, pp. 248–255.

[17] Ahmed M. Daoud , Khlid M. Hosny , Ehab R. Mohamed, Building a New Semantic Social Network Using Semantic Web-Based Techniques, Fusion: Practice and Applications, Vol. 3 , No. 2 , (2021) : 54-65 (Doi : https://doi.org/10.54216/FPA.030201)

[18] K. R. Mehta, A. Deshpande, and S. N. Varma, "Support vector machine–based medical diagnosis using enhanced kernel functions," in Proc. IEEE Int. Conf. Artificial Intelligence in Healthcare (AIH), Singapore, 2022, pp. 89–96.

[19] Ramgude AkshayDili , K. Vengatesa , Kunal Joshi , Chaitanya Tekane, Counterfeit Product Detection System Using One-Time QR code, Journal of Cognitive Human-Computer Interaction, Vol. 2 , No. 2 , (2022) : 65-71 (Doi : https://doi.org/10.54216/JCHCI.020205)

[20] K.Vengatesan , Raghvendra Vijay Naidu , Kunal Ganesh Joshi , ChaitanyaSantosh Tekane , Siddhant Ravindra Gore, Machine Learning Based Product Price Inference Using Price Elasticity of Demand Approach, Journal of Cognitive Human-Computer Interaction, Vol. 3 , No. 1 , (2022) : 08-16 (Doi : https://doi.org/10.54216/JCHCI.030101)

[21] A. U. Haq, J. Li, Z. Ali, M. H. Memon, M. Abbas, and S. Nazir, ''Recognition of the Parkinson's disease using a hybrid feature selection approach,'' J. Intell. Fuzzy Syst., vol. 39, pp. 1–21, Jan. 2020.

[22] A. U. Haq, J. P. Li, J. Khan, M. H. Memon, S. Nazir, S. Ahmad, G. A. Khan, and A. Ali, ''Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data,'' Sensors, vol. 20, no. 9, p. 2649, May 2020.

[23] J. Dheeba, N. A. Singh, and S. T. Selvi, ''Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach,'' J. Biomed. Informat., vol. 49, pp. 45–52, Jun. 2014.

[24] A. Sharma, R. Gupta, and N. Verma, "Decision tree–based predictive modeling for healthcare diagnosis using structured clinical datasets," in Proc. 11th Int. Conf. Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2022, pp. 1–6.

[25] Y. Prasad, K. K. Biswas, and C. K. Jain, ''SVM classifier based feature selection using GA, ACO and PSO for siRNA design,'' in Proc. Int. Conf. Swarm Intell. Berlin, Germany: Springer, 2010, pp. 307–314.

[26] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, ''Breast cancer diagnosis using an unsupervised feature extraction algorithm based on deep learning,'' in Proc. 37th Chin. Control Conf. (CCC), Jul. 2018, pp. 9428–9433.

[27] M. M. Islam, H. Iqbal, M. R. Haque, and M. K. Hasan, ''Prediction of breast cancer using support vector machine and K-nearest neighbors,'' in Proc. IEEE Region Humanitarian Technol. Conf. (R-HTC), Dec. 2017, pp. 226–229.

[28] A. M. Ahmad, G. M. Khan, S. A. Mahmud, and J. F. Miller, ‘‘Breast cancer detection using Cartesian genetic programming evolved artificial neural networks,’’ in Proc. 14th Int. Conf. Genetic Evol. Comput. Conf. (GECCO), New York, NY, USA, 2012, pp. 1031–1038.