The Scientific Temper (2024) Vol. 15 (4): 3187-3195

E-ISSN: 2231-6396, ISSN: 0976-8653

Doi: 10.58414/SCIENTIFICTEMPER.2024.15.4.28

https://scientifictemper.com/

RESEARCH ARTICLE

Improving classification precision for medical decision systems through big data analytics application

Archana G*, Vijayalakshmi V

Abstract

The rapid evolution of machine learning (ML) and big data analytics has modernized medical decision-making procedure, offering promising path for improving classification precision and ultimately, patient outcomes. This research inspects methodologies for enhancing the classification accuracy of medical decision systems by leveraging ML algorithms and big data analytics procedure. In this study, a broad evaluation of existing literature on ML applications in healthcare and medical decision-making is carried out to discover current challenges and potential areas for improvement. The research explores the integration of diverse data sources, including electronic health records (EHRs), medical imaging, genomic data, and patient-generated data, to build robust predictive models. Moreover, the research emphasizes the importance of interpretability and transparency in ML models for medical decision-making, particularly in critical healthcare settings where the rationale behind predictions is crucial. Techniques for model explainability, such as feature importance analysis and model-agnostic interpretability methods, are explored to enhance trust and adoption of ML-driven decision systems by healthcare professionals. Furthermore, the study investigates advanced ML algorithms such as deep learning, ensemble methods, and feature engineering techniques to extract meaningful patterns from large and complex medical datasets. Through experimentation with real-world medical datasets, the efficacy of these algorithms in improving classification accuracy is evaluated and compared against traditional methods. The result of this research contributes to the advancement of ML-driven medical decision systems by providing insights into strategies for improving classification accuracy, thereby facilitating more exact diagnosis, prognosis, and treatment recommendations. Ultimately, the integration of ML and big data analytics holds immense potential for revolutionizing healthcare delivery and improving patient outcomes.

Keywords: Medical decision systems, Big data analytics, Healthcare data, Machine learning, Classification accuracy.

Introduction

In recent years, the proliferation of healthcare data, coupled with advancements in machine learning (ML) and big data analytics, has catalyzed a paradigm shift in medical decision-making progression. Conventional approaches to diagnosis and treatment planning are increasingly augmented, if not supplanted, by sophisticated computational models capable of extracting actionable insights from vast and heterogeneous datasets. In this context, the quest for enhancing classification accuracy in medical decision systems has emerged as a paramount objective, with profound implications for patient care, resource allocation, and healthcare outcomes. (Andreu *et al.,* 2015)

Medical decision systems play a pivotal role in clinical practice, aiding healthcare professionals in diagnosing diseases, predicting patient outcomes, and personalizing treatment regimens. However, the complexity and variability inherent in clinical data pose formidable challenges to traditional analytical techniques. Conventional statistical methods often struggle to capture the intricate relationships within multidimensional datasets comprising diverse modalities such as electronic health records (EHRs), medical imaging, genomic information, and real-time patient-generated data. As a result, there is a growing imperative

PG & Research Department of Computer Science, Government Arts College (Grade-I), (Affiliated to Bharathidasan University) Ariyalur, Tamil Nadu, India.

***Corresponding Author:** Archana G, PG & Research Department of Computer Science, Government Arts College (Grade-I), (Affiliated to Bharathidasan University) Ariyalur, Tamil Nadu, India, E-Mail: archana.scasca@gmail.com

How to cite this article: Archana, G., Vijayalakshmi, V. (2024). Improving classification precision for medical decision systems through big data analytics application. The Scientific Temper, **15**(4):3187-3195.

Doi: 10.58414/SCIENTIFICTEMPER.2024.15.4.28

Source of support: Nil

Conflict of interest: None.

Received: 10/10/2024 **Accepted:** 16/11/2024 **Published** : 20/12/2024

to harness the power of ML and big data analytics to unlock the latent predictive potential of these data sources (Belle *et al.,* 2015).

Machine learning techniques offer a promising avenue for improving classification accuracy in medical decision systems by automatically learning from data patterns and iteratively refining predictive models. Unlike rule-based algorithms, ML algorithms have the capacity to discern intricate patterns, adapt to evolving data distributions, and uncover subtle correlations that may elude human intuition. Moreover, the advent of deep learning architectures has revolutionized the analysis of complex medical images and unstructured clinical text, enabling the development of more accurate and scalable predictive models.

In parallel, big data analytics frameworks provide the infrastructure necessary for processing, storing, and analyzing massive volumes of healthcare data at scale. By leveraging distributed computing platforms and advanced data management techniques, big data analytics enable the integration of disparate data sources and the extraction of actionable insights in near real-time. Consequently, healthcare organizations are empowered to derive valuable clinical insights, optimize resource utilization, and enhance patient outcomes through evidence-based decision-making (Capobianco, 2017) (Table 1).

Against this backdrop, this research aims to explore novel methodologies for enhancing classification accuracy in medical decision systems through the synergistic integration of machine learning and big data analytics. By elucidating the challenges, opportunities, and best practices in this domain, this study seeks to advance the state-of-the-art in ML-driven healthcare analytics and pave the way for more effective, efficient, and patient-centric healthcare delivery (Cunha *et al.,* 2015).

Fuzzy K-Medoids Clustering-Based Attribute Weighting (FKMAW)

The following section outlines the initial steps of fuzzy k-medoid clustering. Let A = {a1, a2, a3, ..., am} represent a collection of m objects, each potentially characterized by a feature vector. Define the dissimilarity between objects as d(aj, ai). Let M be a subset of A with cardinality k, indicating that M is a k-subset of A. The objective of the fuzzy k-medoids algorithm is to minimize (Silva *et al.,* 2015).

$$J(M;A)=\sum_{i=1}^{m}\sum_{j=1}^{k}c_{ji}^{m}d(a_i,m_j)$$

The minimization process is carried out for each M within Ak. Within the equation, cjipredominantly denotes the fuzzy membership of ai within cluster j. The heuristic depiction of the fuzzy membership function is outlined through the subsequent equation:

$$c_{ji}=\frac{\left(\frac{1}{d(a_i,m_j)}\right)^{1/(f-1)}}{\sum_{p=1}^{k}\left(\frac{1}{d(a_i,m_j)}\right)^{1/(f-1)}}$$

$$c_{ji}^{m}=\sum_{p=1}^{k}\frac{\exp\{-\alpha d(a_i,m_j)\}}{\exp\{-\alpha d(a_i,m_j)\}}$$

Equation 2 introduces the concept of a fuzzifier denoted as 'f'. Utilizing the equations provided, a fuzzy partition of set A is established, ensuring that the cumulative membership of an object 'ai' across all classes equals one.

$$c_{ji}=\exp\left(-\frac{d(a_i,m_j)}{\mu_j}\right)$$

Scale parameters α and μ, in equation 4, vary for each cluster based on cluster size, determined from input data. Crucial for identifying cluster boundaries and removing outliers, they convert input features to fuzzy values in the fuzzy k-medoids algorithm. An effective FKMAW method enhances classification accuracy in heart disease datasets by making data linearly separable, employing clustering techniques (Ding *et al.,* 2018).

$$\min Y_{fuskm}=\sum_{a=1}^{n}\sum_{b=1}^{n}d_{ab}(m_{ab})^{p}$$

$$\text{Substitute, } \sum_{b=1}^{n}m_{ab}=1,\quad \forall a\in\{1,2,\ldots,n\}$$

$$m_{ab}\le m_{bb},\quad \forall a,b\in\{1,2,\ldots,n\}$$

$$\sum_{b=1}^{n}e_{bb}=c$$

$$m_{ab}\in[0,1],\ \forall a,b\in\{1,2,\ldots,n\},a\neq b$$

$$e_{bb}\in[0,1],\ \forall b\in\{1,2,\ldots,n\}$$

The dissimilarity measure (d ab) compares the distance between data objects Oa and Ob. Using a fuzzy factor p and membership degree mab, the medoid's presence in a cluster is assessed. A hyperparameter, fuzziness factor, determines cluster overlap. Varying p and mab generates different partitioning strategies, allocating objects accordingly (Farid *et al.,* 2016)

$$m_{ab} = \frac{1}{\sum_{x|m_{xx}=1} \left(\frac{d_{ab}}{d_{bx}}\right)^{1/(p-1)}}$$

FKMAW - Fuzzy k-medoids Clustering based Attribute Weighting

Initially, the fuzzy k-medoids technique identifies cluster medoids. Crucially, two key ratios are derived: medoid or mean value and mean or medoid value. Data values are multiplied by the appropriate ratio, determined by comparison with the medoid or mean. FKMAW ensures medoid proximity (Hernandez *et al.,* 2017).

$$t_1 = \frac{\sum_{h=1}^{s} y_h}{s}$$

$$\omega_1 = \frac{t_1}{x_1} OR\ \omega_1 = \frac{l_1}{t_1} OR\ \omega_1 = 1$$

Here the values of the feature(t_1) are referred to as h=1,2,…, s. The number of associated feature values in c_1 class is represented as s and the average feature value as t_1, respectively. Depending upon the clustering method of k-medoids, the c_1 class with the medoid value for the associated feature is referred to as . A weighting coefficient (ω_1) with a higher value is used when the medoid value computed is found to be higher than the feature value. (Istephan *et al.,* 2015)

Deep Belief Networks Architecture

Understanding deep belief networks (DBN) architecture becomes accessible through RBMs. In thisarchitecture, the initial RBM, equipped with a visible network layer, ingests input data. The subsequent RBM receives input from the output of the preceding RBM. The DBN learning process adheres strictly to a layer-by-layer, unsupervised learning approach. Within DBN architecture, pre-training and fine-tuning emerge as pivotal stages. During pre-training, RBMs undergo continuous training until the hidden layer of the final RBM is reached. Subsequently, in the fine-tuning phase, sample data output labels come into play, facilitating parameter adjustment through the back-propagation algorithm.(Knoppers *et al.,* 2017)

Demonstrating the efficacy of the MDSS, a DBN prediction model is constructed utilizing an extreme learning machine (ELM) (Figure 1). Within this model, the ELM functions as the foundation for regression tasks, showcasing its versatility. The activation function, exemplified by a sigmoid function from the ELM, plays a crucial role in model operations. However, it's noteworthy that the DBN structure is composed of two RBMs. Each layer within the DBN encompasses varying neuron counts, specifically 35, 100, and 10 neurons, contributing to its complexity. A visual representation depicting the architecture of the DBN model is provided (Lo *et al.,* 2016).

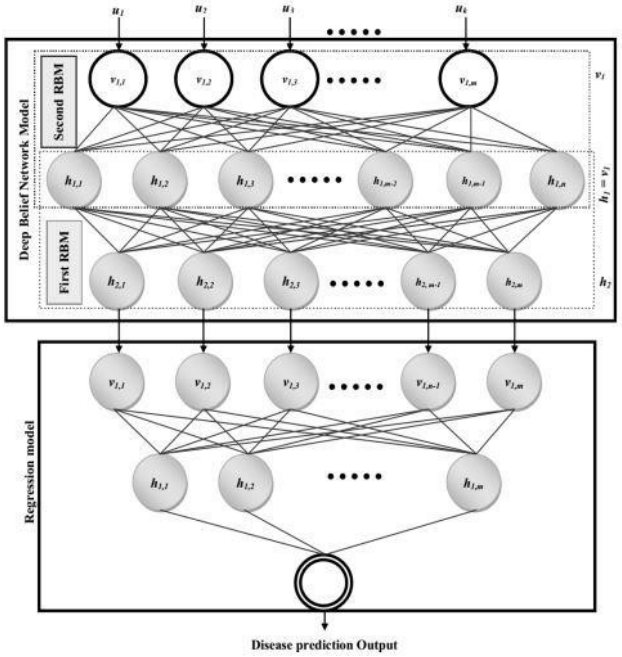

Figure 1: Multi-dimension scaling for disease prediction using DBN with a regression model structure of the proposed DBN-based model

Experimental Results and Analysis

During the analysis, the proposed approach's effectiveness was evaluated across six distinct datasets: heart disease (Hungarian), heart disease (Swiss), heart disease (Cleveland), heart disease (Statlog), BUPA liver disorders, Parkinson's disease (PD), HSV, and early stage diabetes Risk Prediction (ESDRP) dataset. Each dataset is briefly described below:

CAD-based Datasets (Heart Hungary, Heart Swiss, Heart Cleveland)

Obtained from the UCI data repository, these datasets are utilized in CAD analysis. They comprise 13 subsets of specific non-category attributes and 76 category attributes each (Mathew *et al.,* 2015).

Parkinson's Disease (PD) Dataset

This dataset comprises 195 biomedical sound measurements collected from 23 patients with Parkinson's disease and eight patients without the disease.

Statlog Heart Disease Dataset

These datasets, sourced from the UCI repository, include 270 instances from patients affected by heart disease and those without the disease. Among these, 150 samples are from patients with heart disease, while 120 samples are from symptom-free individuals (Mendelson, 2017) (Table 2).

The liver disorders dataset consists of 345 instances categorized into two classes and comprising six attributes.

Table 1: Attributes of the probability distribution dataset

Parameter	*Value*
Sound fundamental frequency	Average, minimum and maximum forms
Measuring disorders based on thefundamental frequency	Jitter: DDP(Dysphonia Detection), Jitter: PPQ(Pitch Perturbation Quotient), Jitter: RAP(Relative Average Perturbation), Jitter (absolute), Jitter (%)
Amplitude irregularity measurements	Shimmer(dB), Shimmer: DDA(Discrete Dipole Approximation), APQ5(Amplitude Perturbation Quotient), APQ3, APQ,
The ratio between noise and audio tonecomponent	HNR(Harmonic-to-Noise Ratio) and NHR(Noise –to-Harmonic Ratio)
Nonlinear dynamic complexity dimensions	RPDE (Recurrence Period Density Entropy) and D2(Correlation Dimension)
Three different trials of Fundamentalfrequency variation	spreads 1, 2 and PPE(Pitch Period Entropy)
spreads 1, 2 and PPE (Personal Protective Equipment)	DFA(Defense Production Act)

This dataset was compiled by the scientific research company BUPA (Table 3 and Figure 2). Data samples were collected from patients with hepatic deficiency as well as from healthy individuals without any disability. Specifically, there are 200 samples from healthy individuals and 145 samples from patients with hepatic disability. Each sample is characterized by six major features, with the first five features derived from the patient's blood test report. The final feature is contingent upon the individual's alcohol intake (Murphy *et al.,* 2017).

Cardiovascular Disease Dataset

The cardiovascular disease dataset includes three distinct types of input features: objective, examination, and subjective. Objective data contains factual information, while the examination feature encompasses medical examination results. Information provided by the patient is recorded in the subjective feature. Table 2 outlines the features present in the cardiovascular disease dataset (Ni *et al.,* 2015).

Physionet Heart Disease Dataset

The dataset comprises patient health records collected as part of routine procedures. These records were gathered during mandatory follow-up appointments at 28 days, 3 months,and 6 months intervals. In cases where patients couldn't physically visit the clinic, follow-up was conducted via telephone. Patients admitted to Zigong Fourth People's Hospital with heart disease between December 2016 and June 2019 were retrospectively included in the study. Due to its retrospective nature, informed consent was waived. The analysis adhered to the principles of the Helsinki Declaration and was approved by the Zigong Fourth People's Hospital Ethics Committee (approval number 2020-010). Electronic health records of successive patients diagnosed with heart disease were examined, with heart failure diagnosed based on European Society of Cardiology (ESC) standards (Olaronke *et al.,* 2016).

- Heart disease symptoms and/or signs may manifest as breathlessness, orthopnea, nocturnal dyspnea paroxysmal, decreased resistance to exercises, nausea, fatigue, prolonged healing time following exercise, and swelling of the ankle.
- Common signs include elevated jugular venous pressure, hepatojugular reflux, presence of a third heart tone (gallop rhythm), and lateral displacement of the apical impulse.
- Elevated brain natriuretic peptide levels (BNP) (BNP > 35 pg/mL and/or NT- proBNP > 125 pg/mL) are indicative.
- Objective confirmation of other underlying cardiac functional and systemic improvements in heart failure is essential.

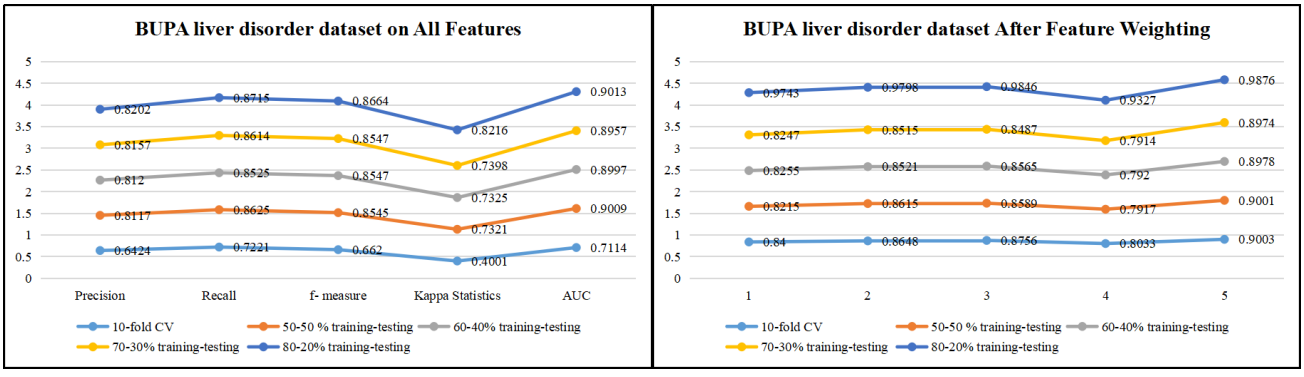

Figure 2: Graphs for results acquired with BUPA liver disorder dataset on various performance indicators

Table 2: Features present in the cardiovascular disease dataset

S. No	*Name*	*Value*
Objective features		
1	Age	Integer(days)
2	Height	Integer(cm)
3	Weight	Float(kg)
4	Gender	Categorical Code
Examination feature		
5	Systolic Blood Pressure	Integer(ap_hi)
6	Diastolic Blood Pressure	Integer(ap_lo)
7	Cholesterol	Normal Above Normal Well Above Normal
8	Glucose	Normal Above Normal Well Above Normal
Subjective feature		
9	Smoking	Smoke(Binary)
10	Alcohol Intake	Alco(Binary)
11	Physical Activity	Active(Binary)
12	Presence or Absence of cardiac disease	Target Variable(Binary)

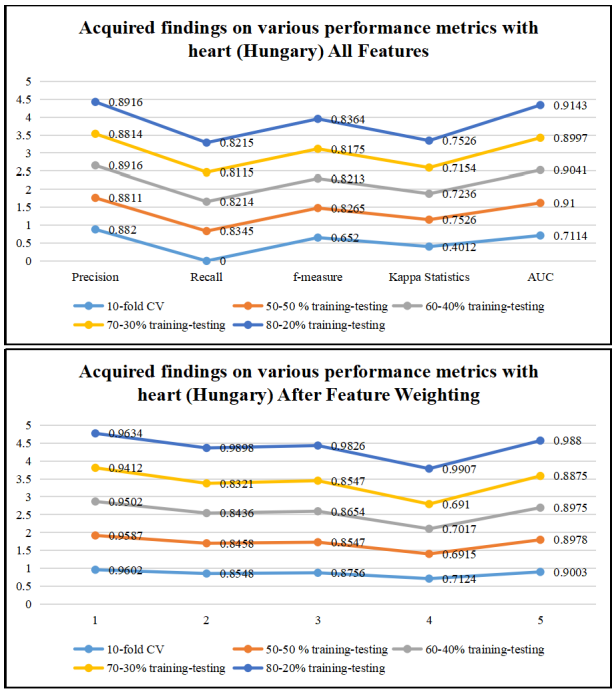

Figure 3: Graphs for acquired findings on various performance metrics with heart (Hungary) dataset

- In cases of confusion, a stress test or an invasively assessed elevated left ventricular (LV) filling pressure may be necessary to validate the diagnosis.

The Parkinson's disease dataset findings are summarized in the table. Using the fuzzy k-medoids attribute weighting (FKMAW) + deep belief network- extreme learning machine (DBNKELM) method with 10-fold cross-validation (CV), a precision rate of classification at 0.9487 is achieved. Comparatively, the DBNKELM algorithm applied to the original dataset yields a precision rate of 0.6353. In contrast, the FKMAW + DBNKELM method attains a higher precision rate of 0.9745 with an 80-20 % training-testing ratio. Conversely, combining the original dataset with the DBNKELEM algorithm results in a precision classification rate of 0.6460. Notably, the FKMAW + DBNKELM technique exhibits superior performance in terms of kappa statistic value. Specifically, the 10-fold CV system achieves the highest kappa measure of 0.9818. This disparity highlights potential variations between the two approaches. Lastly, the PD dataset utilizes weighted features to derive effective results (Özdemir *et al.*, 2018).

Table 4 presents the results obtained with the heart (Hungary) dataset (Figure 3). Additionally, the FKMAW +

Table 3: Results acquired with BUPA liver disorder dataset on various performanceindicators

Features	*All Features*				
Metrics	*Precision*	*Recall*	*F-measure*	*Kappa statistics*	*AUC*
10-fold CV	0.8614	0.7410	0.7581	0.6732	0.5431
50–50 % training-testing	0.8532	0.6975	0.7389	0.6247	0.5689
60–40% training-testing	0.8314	0.6854	0.7265	0.6104	0.5432
70–30% training-testing	0.8311	0.6935	0.7187	0.6231	0.5278
80–20% training-testing	0.8615	0.7092	0.7486	0.6487	0.5842
After Feature Weighting					
10–fold CV	0.9202	0.8821	0.9221	0.9205	0.8342
50–50 % training-testing	0.9104	0.8715	0.9147	0.9105	8278
60–40% training-testing	0.9003	0.8614	0.9139	0.9009	8189
70–30% training-testing	0.9104	0.8715	0.9157	0.9147	8200
80–20% training-testing	0.9534	0.8984	0.9247	0.9282	0.8118

Table 4: Acquired findings on various performance metrics with heart (Hungary) dataset

Features	*All features*				
Metrics	*Precision*	*Recall*	*f-measure*	*Kappa Statistics*	*AUC*
10-foldCV	0.8820	0.70.21	0.652	0.4012	0.7114
50–50%training-testing	0.8811	0.8345	0.8265	0.7526	0.9100
60–40%training-testing	0.8916	0.8214	0.8213	0.7236	0.9041
70–30%training-testing	0.8814	0.8115	0.8175	0.7154	0.8997
80–20%training-testing	0.8916	0.8215	0.8364	0.7526	0.9143
After Feature Weighting					
10-foldCV	0.9602	0.8548	0.8756	0.7124	0.9003
50–50%training-testing	0.9587	0.8458	0.8547	0.6915	0.8978
60–40%training-testing	0.9502	0.8436	0.8654	0.7017	0.8975
70–30%training-testing	0.9412	0.8321	0.8547	0.6910	0.8875
80–20%training-testing	0.9634	0.9898	0.9826	0.9907	0.9880

Table 5: The effects of the heart (Swiss) dataset on different performance evaluation criteria

Features	*All features*				
Metrics	*Precision*	*Recall*	*f-measure*	*Kappa statistics*	*AUC*
10-foldCV	0.9020	0.7021	0.632	0.4005	0.7014
50–50%training-testing	0.8942	0.8321	0.8541	0.7547	0.9213
60–40%training-testing	0.8936	0.8317	0.8498	0.7478	0.9233
70–30%training-testing	0.8816	0.8287	0.8325	0.7324	0.9185
80–20%training-testing	0.9053	0.8415	0.8664	0.7616	0.9153
After Feature Weighting					
10-foldCV	0.9262	0.9398	0.9756	0.7817	0.9898
50–50%training-testing	0.9512	0.8432	0.8547	0.6924	0.9021
60–40%training-testing	0.9407	0.8327	0.8432	0.6814	0.8975
70–30%training-testing	0.9315	0.8215	0.8412	0.6809	0.8950
80–20%training-testing	0.9602	0.8548	0.8646	0.8234	0.9104

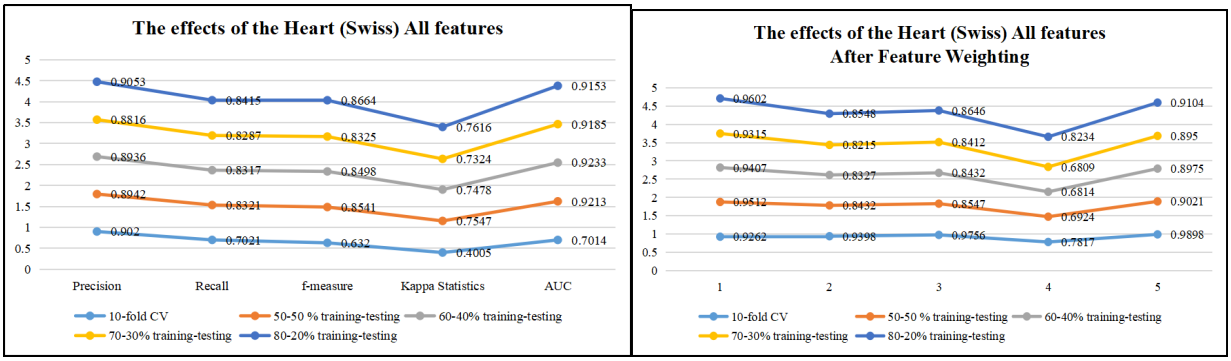

Figure 4: Graphs for the effects of the heart (Swiss) dataset on different performance evaluation criteria

DBNKELM classification precision rate using the 10-fold CV technique is 0.9602. Comparatively, the precision rate obtained from the original dataset using the DBNKELM algorithm is 0.8820. Furthermore, the FKMAW + DBNKELM method achieves a precision rate of 0.9634 with an 80 to 20% training-testing split. Conversely, integrating the actual

Table 6: Results from heart (Cleveland) data set on different performance evaluation criteria

Features	*All features*				
Metrics	*Precision*	*Recall*	*f-measure*	*KappaStatistics*	*AUC*
10-foldCV	0.8932	0.7220	0.8156	0.6654	0.7614
50–50%training-testing	0.8857	0.8039	0.8314	0.7100	0.8975
60–40%training-testing	0.8762	0.7932	0.8214	0.6912	0.8947
70–30%training-testing	0.8721	0.7923	0.8145	0.6813	0.8921
80–20%training-testing	0.8962	0.8115	0.8164	0.7716	0.9001
After Feature Weighting					
10-foldCV	0.9653	0.8048	0.8546	0.7124	0.9123
50–50%training-testing	0.9544	0.8021	0.8025	0.7766	0.9123
60–40%training-testing	0.9432	0.7944	0.8365	0.7110	0.9045
70–30%training-testing	0.9317	0.7932	0.8245	0.7009	0.8978
80–20%training-testing	0.**9762**	0.**9198**	0.9576	0.9617	0.9850

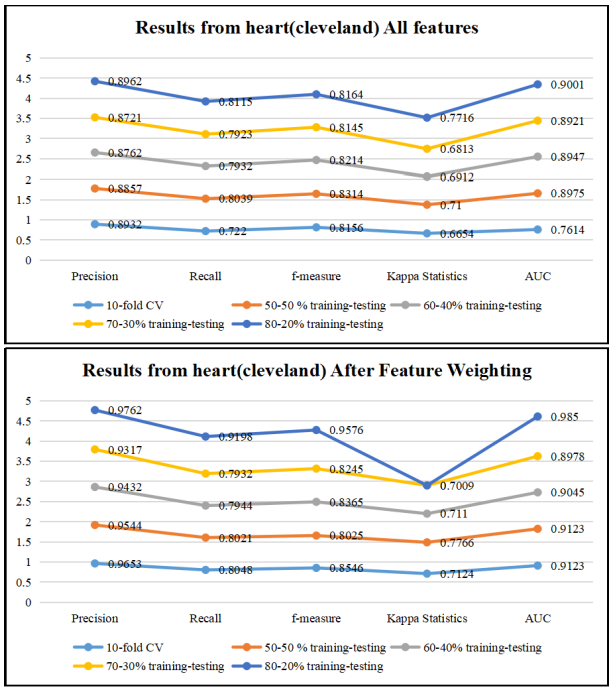

Figure 5: Graphs for results from heart (Cleveland) data set on different performance evaluation criteria

dataset with the DBNKELEM classifier (i.e., the actual dataset + classifier) yields a recall classification rate of 0.8916. Notably, the FKMAW + DBNKELM approach demonstrates superior results, particularly in terms of kappa statistics significance. Despite this, the highest kappa value, obtained by FKMAW + DBNKELM with the 80 to 20% training- testing technique, is 0.9907. Ultimately, the heart (Hungary) dataset leverages positively weighted features to enhance outcomes (Panda *et al.,* 2017).

Table 5 summarizes the findings obtained with the Heart (Swiss) dataset (Figure 4). The FKMAW + DBNKELM classification precision rate using the 10-fold cross-validation (CV) approach is 0.9262. In comparison, the precision rate obtained from the initial dataset using the DBNKELM algorithm with 10-fold CV is 0.9020. Additionally, the FKMAW + DBNKELM method achieves a precision rate of 0.9602 with an 80 to 20% training-testing split. Conversely, integrating the actual dataset with the DBNKELEM classifier results in a precision classification rate of 0.9053. Notably, superior results are obtained with the FKMAW + DBNKELM approach, particularly when considering the kappa statistic value (0.8234) for an80-20% training-testing split (Pramanik *et al.,* 2017).

Table 6 displays the findings obtained from the heart (Cleveland) dataset (Figure 5). The classification precision rate using the 10-fold CV process for FKMAW + DBNKELM is 0.9653, while using the original dataset with the DBNKELM algorithm yields a precision rateof 0.8932. Conversely, the FKMAW + DBNKELM technique achieves a precision rate of 0.9762 with an 80 to 20% training-testing split. However, merging the original dataset with the DBNKELEM classifier results in a precision classification rate of 0.8962. Notably, theFKMAW + DBNKELM approach shows improved results, particularly in terms of kappa statistics value. With the 80 to 20% training-testing split technique, FKMAW + DBNKELM achieves the highest kappa score of 0.9850. Additionally, the heart (Cleveland) dataset is better at acquiring positively weighted features (Price *et al.,* 2015).

Furthermore, Table 7 presents the findings obtained with the dataset for cardiovascular diseases. The precise classification rate using the 10-fold CV method for FKMAW + DBNKELM is 0.9354. In contrast, FKMAW + DBNKELM achieves a precision rate of 0.9765 for an 80 to 20% training-testing split before attribute weighting. However, when integrating the actual dataset with the DBNKELEM classifier (i.e., the actual dataset +classifier), the precise classification rate achieved is 0.9867. Notably, the FKMAW + DBNKELM approach demonstrates superior results, especially when

Table 7: Cardiovascular disease dataset results on different performance metrics

Features	*All features*				
Metrics	*Precision*	*Recall*	*f-measure*	*Kappa Statistics*	*AUC*
10-foldCV	0.9235	0.8520	0.8657	0.8547	0.7944
50–50%training-testing	0.9737	0.8854	0.8187	0.8214	0.9358
60–40%training-testing	0.9754	0.8843	0.8298	0.8179	0.9547
70–30%training-testing	0.9745	0.8857	0.8321	0.8187	0.9598
80–20%training-testing	0.9765	0.8969	0.8451	0.8257	0.9695
After Feature Weighting					
10-foldCV	0.9354	0.8548	0.9589	0.8045	0.9525
50–50%training-testing	0.9632	0.8739	0.8978	0.7945	0.9845
60–40%training-testing	0.9539	0.8640	0.8954	0.7936	0.9789
70–30%training-testing	0.9547	0.8628	0.8735	0.7923	0.9754
80–20%training-testing	0.9867	0.9694	0.9984	0.9561	0.9894

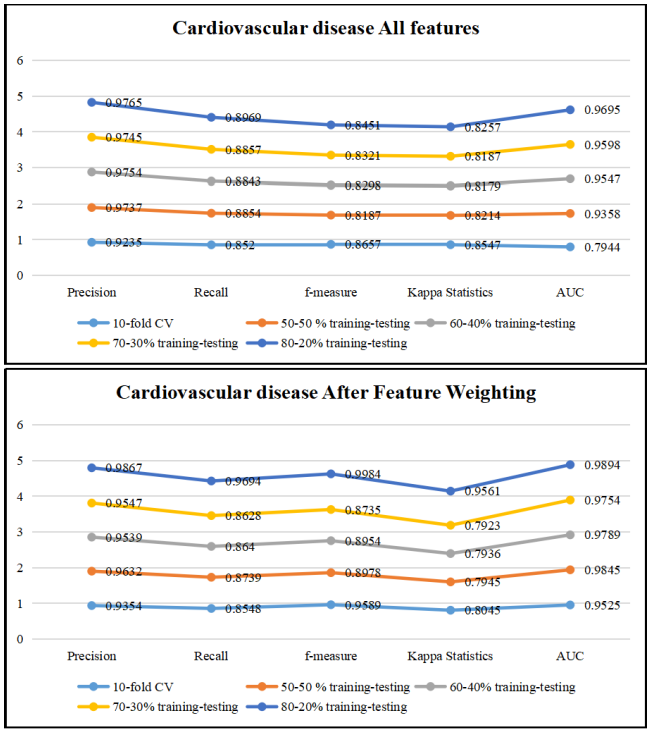

Figure 6: Graphs for Cardiovascular disease dataset results on different performance metrics

considering the significance of kappa statistics. Moreover, with an 80 to 20% training-testing split, FKMAW + DBNKELM attains the highest kappa value of 0.9561. Ultimately, the cardiovascular disease dataset emphasizes positively weighted features (Reddy *et al.,* 2016).

Table presents the results obtained from the physionet heart disease dataset. The classification precision rate of FKMAW + DBNKELM using the 10-fold cross-validation (CV) process is 0.9872. Conversely, FKMAW + DBNKELM achieve a precision rate of 0.9987 with an 80 to 20% training-testing split. Additionally, integrating the original dataset with the DBNKELEM algorithm (i.e., the original dataset + DBNKELM) yields a precision classification rate of 0.9562. Notably, the FKMAW + DBNKELM approach demonstrates positive performance, particularly in terms of the kappa statistics value. However, with the 80 to 20% training-testing phase, the FKMAW + DBNKELM method attains the highest kappa measure of 0.9654. Ultimately, the physionet heart disease dataset is adept at acquiring positively weighted characteristics (Ren *et al.,* 2015).

Conclusion

Conclusively, the utilization of big data analytics presents significant potential for enhancing the precision of medical decision systems. Through the analysis of extensive healthcare data, these systems are better equipped to recognize essential patterns, trends, and correlations necessary for accurate classification. By employing advanced analytics techniques, healthcare professionals can make more informed decisions, leading to improved patient outcomes and the progression of medical understanding. Thus, the integration of big data analytics into medical decision systems marks a crucial advancement towards attaining heightened levels of classification accuracy within healthcare environments.

Acknowledgment

I am extremely grateful to my supervisors, Dr. V. Vijayalakshmi, for her valuable advice, continuous support, and patience during my study. Her immense knowledge and plentiful experience have encouraged me in all the time of my academic research.

References

Andreu-Perez, J., Poon, C. C., Merrifield, R. D., Wong, S. T., & Yang, G. Z. (2015). Bigdata for health. IEEE J Biomed Health Inform, 19(4), 1193-1208.

Belle, A., Thiagarajan, R., Soroushmehr, S. M., Navidi, F., Beard, D. A.,

& Najarian, K.(2015). Big data analytics in healthcare. BioMed Research International, 2015.

Capobianco, E. (2017). Systems and precision medicine approaches to diabetes heterogeneity: a Big Data perspective. Clinical and Translational Medicine, 6(1), 23.

Cunha, J., Silva, C., & Antunes, M. (2015). Health twitter big bata management withhadoop framework. Procedia Computer Science, 64, 425-431.

De Silva, D., Burstein, F., Jelinek, H. F., & Stranieri, A. (2015). Addressing the complexities of big data analytics in healthcare: the diabetes screening case. AustralasianJournal of Information Systems, 19, S99-S115.

Ding, W., Lin, C. T., Chen, S., Zhang, X., & Hu, B. (2018). Multiagent-consensus- MapReduce-based attribute reduction using co-evolutionary quantum PSO for big data applications. Neurocomputing, 272, 136-153.

Farid, D. M., Nowe, A., & Manderick, B. (2016, December). A feature grouping method for ensemble clustering of high-dimensional genomic big data. In Future Technologies Conference (FTC) (pp. 260-268). IEEE.

Hernandez, I., & Zhang, Y. (2017). Using predictive analytics and big data to optimize pharmaceutical outcomes. American Journal of Health-System Pharmacy, 74(18), 1494- 1500.

Istephan, S., & Siadat, M. R. (2015, November). Extensible query framework for unstructured medical data-a big data approach. In Data Mining Workshop (ICDMW), 2015 IEEE International Conference on (pp. 455-462). IEEE.

Knoppers, B. M., & Thorogood, A. M. (2017). Ethics and big data in health. Current Opinion in Systems Biology, 4, 53-57.

Lo'ai, A. T., Mehmood, R., Benkhlifa, E., & Song, H. (2016). Mobile cloud computing model and big data analysis for healthcare applications. IEEE Access, 4, 6171-6180.

Mathew, P. S., & Pillai, A. S. (2015, March). Big data solutions in healthcare: problems and perspectives. In Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference on (pp. 1-6). IEEE.

Mohamed Azharudheen A, Dr.Vijayalakshmi V,(2024) Improvement of Data Analysis and Protection using Novel Privacy-Preserving Methods for big data application, The Scientific Temper (2024) Vol. 15 (2): 2181-2189, E-ISSN: 2231-6396, ISSN: 0976-8653.

Mohamed Azharudheen A, Dr.Vijayalakshmi V,(2024) Analyze the New Data Protection Mechanism to Maximize Data Availability without Having Compromise Data Privacy, Educational Administration: Theory and Practice 2024, Vol. 30(5), 3911-3922, ISSN: 2148-2403.

Murphy, S. N., Avillach, P., Bellazzi, R., Phillips, L., Gabetta, M., Eran, A., & Kohane, I.S. (2017). Combining clinical and genomics queries using i2b2– Three methods. PloS one,12(4), e0172187.

Ni, J., Chen, Y., Sha, J., & Zhang, M. (2015, November). Hadoop-based distributed computing algorithms for healthcare and clinic data processing. In Internet Computingfor Science and Engineering (ICICSE), 2015 Eighth International Conference on (pp. 188-193). IEEE. International Journal of Mathematical, Engineering and Management Sciences Vol. 4, No. 1, 17–26, 2019 ISSN: 2455-7749

Olaronke, I., & Oluwaseun, O. (2016, December). Big data in healthcare: prospects, challenges and resolutions. In Future Technologies Conference (FTC) (pp. 1152-1157). IEEE.

Özdemir, V., & Hekim, N. (2018). Birth of industry 5.0: Making sense of big data with artificial intelligence,"the internet of things" and next-generation technology policy.Omics: A Journal of Integrative Biology, 22(1), 65-76.

Panda, M., Ali, S. M., & Panda, S. K. (2017, March). Big data in health care: A mobile based solution. In Big Data Analytics and Computational Intelligence (ICBDAC), 2017 International Conference on (pp. 149-152). IEEE.

Pramanik, M. I., Lau, R. Y., Demirkan, H., & Azad, M. A. K. (2017). Smart health: big data enabled health paradigm within smart cities. Expert Systems with Applications, 87, 370-383.

Price, W., & Nicholson, I. I. (2015). Big Data, Patents, and the Future of Medicine. Cardozo Law Review, 37, 1401-1453.

Reddy, A. R., & Kumar, P. S. (2016, February). Predictive big data analytics in healthcare. In Computational Intelligence & Communication Technology (CICT), 2016 Second International Conference on (pp. 623-626). IEEE.

Ren, G., & Krawetz, R. (2015). Applying computation biology and "big data" to develop multiplex diagnostics for complex chronic diseases such as osteoarthritis. Biomarkers, 20(8), 533-539.