International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 05 July 2023 **Revised:** 25 August 2023 **Accepted:** 15 September 2023*

A Robust Intuitionistic Fuzzy Constraint Score based Potential Feature Subset Selection for Chronic Diseases Detection

Ms. S. Sabeena[1], Dr. B. Sarojini[2]
[1]Research Scholar, Department of Computer Science
Avinashilingam Institute for Home Science and Higher Education for Women
Coimbatore, India
sabeena.phd@gmail.com
[2]Assistant Professor (SS), Department of Computer Science
Avinashilingam Institute for Home Science and Higher Education for Women
Coimbatore, India
dr.b.sarojini@gmail.com

Abstract: This work proposes a novel feature selection algorithm for high-dimensional features in real-time datasets for prediction or classification. Conventional methods assume dataset values as crisp formats, but in real datasets, instances are represented in linguistic formats, requiring the use of uncertainty theories. The Intuitionistic Fuzzy Similarity based constraint score is proposed, where each feature is denoted as an independent variable and the class variable as a dependent variable. The features are represented in triplet form, with grade of belongingness, non-belongingness, and hesitancy index to maximize relevancy and reduce redundancy. Pairwise similarity matching is computed using Intuitionistic fuzzy similarity distance measure for supervised learning and intuitionistic fuzzy K-NN for semi-supervised learning. Potential feature subsets are selected and validated using deep learning algorithms. The results show that the proposed Intuitionistic fuzzy Constraint score feature selection algorithm produces optimal results compared to other state-of-the-art methods in chronic disease prediction.

Keywords: Chronic Disease, Feature selection, Intuitionistic fuzzy, Supervised, Semi-Supervised

I. INTRODUCTION

As chronic diseases last a very long period, diagnosing them is crucial in the field of medicine. Cancer, Heart disease, Diabetes, and stroke are the most common chronic diseases. Early chronic disease detection facilitates the implementation of preventative measures, and early diagnosis, and timely treatment for patients has consistently been demonstrated to be beneficial [1]. Maintaining clinical databases is now a vital duty in the healthcare sector. To deliver excellent services, the patient data, which includes multiple traits and symptoms associated with disease, needs to be recorded with the utmost diligence [2]. Medical data mining for disease prediction becomes challenging since medical databases often contain redundant and insufficient information. Before using data mining techniques, it is crucial to have proper data pre-processing and data diminution because it can affect the results of mining [3]. If data is precise, reliable, and noise-free, disease prediction becomes rapid and easier.

The selection of features is crucial in data pre-processing methods for the prediction of chronic disease at its early stage more precisely. The most essential disease-related risk factors must be identified during the diagnosis process. Removing unused, duplicated attributes from the chronic disease dataset with the use of relevant feature identification leads to more rapid and effective results [4]. Identifying a subset of characteristics from the initial collection of features and creating patterns in a dataset are both steps in the process of determining the best feature set for the specified processing aim and criterion. It decreases the number of features, eliminates irrelevant, redundant, or noisy features, and has immediate implications for applications, such as accelerating a data mining algorithm and enhancing mining performance like classification accuracy and results in general [5].

Using feature selection techniques, one can search through feature subsets and attempt to identify the top 2N candidate subset from the competition. Even for a large feature set (N), this process is exhaustive and could be too expensive [6, 7]. Other approaches that rely on intuitive or randomized search strategies make performance sacrifices to reduce the computational cost.

Subset generation is essentially a heuristic search procedure where every state in the area of search specifies a potential subset to be considered [8]. Two fundamental considerations define the purpose of this procedure. First, the search's beginning place, which then determines its course. A forward search begins with an empty set and adds features one at a time, a backward search begins with a full set and subtracts features one at a time, or a bi-directional search begins with both sides and simultaneously adds and subtracts features. Starting with a subset chosen at random is another option occasionally chosen on a search strategy second. There are 2^N potential subsets for an N-feature dataset. At even a moderate N, the resulting search space is progressively too large for a comprehensive search.

In this work, the problem of selecting potential features during the presence of uncertainty due to vagueness and incompleteness in the dataset is handled to improve the accuracy of chronic disease detection. Both the dataset with labelled and unlabelled records are utilized to produce an optimized feature selection algorithm using uncertainty theory. A detailed description of the working principle of the proposed intuitionistic fuzzy constraint score-based feature subset selection is explained in the following sections.

1.1 Organization of the paper

The rest of the paper is organized as follows. Section 2 narrates the related works. Section 3 describes the background study of Intuitionistic Fuzzy Sets. Finally, Section 4 encompasses the efficiency analysis followed by the conclusion.

II. RELATED WORKS

Mishra et al [9] conducted a thorough analysis of the influence of wrapper and filter models on classification techniques. The authors used three different methods of filter approach such as information gain, correlation, and chi-square, and three different wrapper approaches like linear forward, best first, and greedy algorithm. The chronic disease datasets are used for feature selection and the decision tree is used as the validator. The attribute significance is done based on the correlation and the mutual information of the independent and dependent variables, but when there is class imbalance it is not possible to use these standard methods.

Divya et al [10] predicted the presence of chronic disease by devising an adaptive-based classifier. They integrated both the principal component analyzer and the ReliefF method for performing feature selection. An optimized support vector machine is used for evaluating nine different diseases for predicting their presence. But the problem of uncertainty conditions and the unlabelled datasets are not considered in this work.

Hegde et al [11] developed an adaptive model for performing feature selection in chronic disease detection. This method selects the features by applying adaptive probabilistic divergence along with a logistic regression model hyper parameterized. The class imbalance and the vague, inconsistent real-time chronic disease diagnosis are not focused.

Ebiaredoh et al [12] designed a boosting classifier using a cost-sensitive model to detect chronic kidney disease and the feature selection is done using information gain. The reduced feature subset is used for chronic kidney disease detection. But the redundancy due to the vague information is not handled by the information gain selection model which affects the reliability of the prediction process.

Esenogho et al [13] constructed an ensemble neural network classifier to predict credit card fraud using the resampling method of hybrid data. The adaptive boosting is fused in long short-term memory for performing the prediction process. The data resampling is accomplished using the hybrid resampling model. While resampling is done randomly, there is a high probability of missing potential data samples due to class imbalance.

Saranya et al [14] conducted a detailed examination of the feature selection process using machine learning algorithms on different chronic datasets. They stated in their work due to the high degree of dependence among input variables in the high dimensional datasets, it suffers from irrelevant or zero significant features issues while discovering feature selection.

Xie et al [15] designed an improved maximal relevance and minimal redundancy model (ImRMR) which applies both mutual information and Pearson correlation for computing the relevance of a single attribute. The candidate feature subset is ranked based on their relevancy and these two measures are used for determining potential features. For voluminous datasets, this statistical approach will lead to high computational complexity.

Abderezak Salmi et al [16] in their work concentrated on handling the problem of high dimensionality in the prediction process. The common constraint score models which evaluate each feature based on two different constraints such as must link and cannot link estimate the relevance is determined on high dimensionality feature space which has a high chance of corrupting the dimension of the dataset. So, they introduced a novel similarity matrix based on the similarity matrix is computed for feature selection.

The above-discussed feature selection models are focused on the computation of datasets with labelled datasets. The unlabelled datasets are not mostly focused on the existing works, but in real-time the medical datasets are often unlabelled. The class imbalance in the chronic disease dataset also affects the process of feature subsets to determine the potential set of features to improve the efficiency of the prediction process. Hence, in this work, the problem of acquiring information about the features in the unlabelled dataset and the labelled dataset is also used for determining the potential feature subset by devising the intuitionistic fuzzy constraint score.

III. BACKGROUND STUDY

Tuned Relief Feature subset

In this work, initial feature subsets are selected using the tuned relief feature selection algorithm, which is a variant of conventional feature selection. To compute the score of each feature in the chronic dataset relief feature the similarity among the most relevant matching nearest neighbour are found. The feature score will be low when a hit, or variation in feature value, is discovered in a close case pair of an identical class. The feature score increases when a close instance pair of instances with differing values for classes has a "miss," or feature value distinction. The recursive feature exclusion strategy is used in this model. After each iteration, the features with the lowest scores are eliminated in order to be included subsequently in feature weight adjustments and proximity calculations. This method computes the relevance score to strengthen the weak convergence of scoring-based parameter update.

The initial feature set for chronic disease detection is determined by applying the tuned relief feature selection. But, the problem of uncertainty due to presence of vagueness and inconsistent cannot be handled by the tuned relief feature selection, hence in this proposed work initial feature subset are chosen by the turned relief and to handle the uncertainty cases the process of intuitionistic fuzzy similarity score is computed for identifying the potential subset of features. This double selection model produces robust results in determining the chronic disease at its early stage by eliminating the redundant and irrelevant features.

IV. PROPOSED WORK

4.1 Intuitionistic Fuzzy Sets

The intuitionistic fuzzy is the generalization of fuzzy sets, which introduces two additional elements known as non-belongingness and hesitation index [17-19]. The fuzzy uses only the degree of belongingness and the problem of vague and hesitancy is not precisely defined in it. Hence in this proposed work, the concept of intuitionistic fuzzy set is used to define uncertainty factors in discovering the best feature subset to predict chronic diseases at their early stage. The intuitionistic fuzzy set R is denoted in the form of

$$\mathcal{H} =: \{(z, B_{\mathcal{H}}(z), NB_{\mathcal{H}}(z))|z \in R\}$$

The belongingness grade and non-belongingness grade are denoted as $B_{\mathcal{H}}(z)$ and $NB_{\mathcal{H}}(z)$ respectively with the mapping value of $B_{\mathcal{H}}(z), NB_{\mathcal{H}}(z) \rightarrow [0,1]$. The condition $0\leq t\ B_{\mathcal{H}}(z) +\ NB_{\mathcal{H}}(z) \leq 1$ is always true for $z \in R$,. The intuitionistic fuzzy introduces an important factor known as hesitation grade $\varphi_{\mathcal{H}}(\mathrm{z}) = 1 - (B_{\mathcal{H}}(\mathrm{z}) + NB_{\mathcal{H}}(\mathrm{z}))$. Collection of all intuitionistic fuzzy sets is represented as IFZ(R).

Definition 1: The intuitionistic fuzzy union operation is defined as

$$C \cup D = \{(z, \mathrm{Max}(B_C(\mathrm{z}), B_D(\mathrm{z})\}, Min\{NB_C(\mathrm{z}), NB_D(\mathrm{z})\})|\mathrm{z} \in \mathrm{R}$$

Definition 2: The intuitionistic fuzzy intersection operation is defined as

$$C \cap D = \{(z, \mathrm{Min}(B_C(\mathrm{z}), B_D(\mathrm{z})\}, Max\{NB_C(\mathrm{z}), NB_D(\mathrm{z})\})|\mathrm{z} \in \mathrm{R}$$

Definition 3: The intuitionistic fuzzy subset operation is defined as

$$C \subset D \Leftrightarrow \forall z \in R, both\ B_C(\mathrm{z}) \leq B_D(\mathrm{z})\ and\ NB_C(\mathrm{z}) \geq NB_D(\mathrm{z})$$

4.1.1 Feature selection using Intuitionistic Fuzzy Constraint Score

To handle the real-time datasets, which are often in linguistic terms, instead of using direct similarity matrices to determine constraint score, in this proposed work uncertainty based intuitionistic fuzzy similarity matrices is developed to evaluate the relevancy of subset of k features represented as $\mathcal{F}_k = \{\mathcal{F}_1, \mathcal{F}_2, .., \mathcal{F}_k\}$ where k = 1..g. The proposed intuitionistic fuzzy similarity score (is applied for both supervised (SV) and semi-supervised (SSV) feature selection. While performing supervised feature selection only intuitionistic fuzzy pairwise constraints are considered by using must link and cannot link to choose the subset of the feature set ($\mathfrak{T}^{SV}(\mathcal{F}_k)$). Whereas in semi-supervised the intuitionistic fuzzy pairwise and the information obtained from the unlabelled dataset is also used for feature selection $\mathfrak{T}^{SSV}(\mathcal{F}_k)$).

The relevance of the feature set $\mathcal{F}_k$ is computed by the intuitionistic fuzzy distance measure among the target Intuitionistic fuzzy similarity matrix $\breve{\varpi}^*$ evaluated using constraint scores of SS or SSV and a similarity matrix $\omega(\mathcal{F}_k)$ computed with $\mathcal{F}_k$ feature subset. The intuitionistic fuzzy score $\mathfrak{T}^*(\mathcal{F}_k)$ must be as low as possible as mentioned in the below equation

$$\mathfrak{T}^*(\mathcal{F}_k) = \sum_{i=1}^{m}\sum_{j=1}^{m}(\omega_{ij}(\mathcal{F}_k) - \breve{\varpi}_{ij}^*)^2$$

Where $\omega(\mathcal{F}_k) \in D^{mxm}$ is the intuitionistic fuzzy similarity matrix computed on the chronic disease datasets Z with the subset of features $\mathcal{F}_k$

$$\omega_{ij}(\mathcal{F}_k) = \exp\left(-\frac{\tau^2(B_i(z),NB_i(z))^k,(B_j(z),NB_j(z))^k}{2\beta^2}\right) ;\ ,\mathrm{j} = 1,2,3..\mathrm{m}$$

Where $(B_i(z), NB_i(z)$ is the degree of belongingness and non-belongingness of the ith record in the chronic disease dataset, represented with the k feature subset. Where τ is the intuitionistic fuzzy similarity distance measure, β^2 is the scaling parameter

4.1.2 Intuitionistic Fuzzy Pairwise Constraint for both Supervised SS and Semi-Supervised (SSV) Feature Selection

In chronic disease datasets, only limited records of information are available on the training dataset. This prior information is expressed by only a few labeled records. It can also be expressed by pairwise constraint that mentions if data samples belong to the same class (ML) or to different classes (CL). The intuitionistic fuzzy pairwise constraint can be given by the user or generated from labeled data samples.

This work considers only a few labeled data samples known as prototypes that characterize the ‘v’ classes. Let $\mathrm{Z}^1 \in D^{p*g}$, $(Z^l \subset Z)$ be the set of ‘p’ prototypes that are associated with class γ^{l}. From the overall set of prototypes that are associated with class γ^{l} $(Z^l \cup_{t=1,..,v} Z^{l,}\mathrm{l}\ =\ 1,2,..\mathrm{v})$, with this construct set must link constraint ML of $(v.p.(p-1))\,p$ airs of must link that are composed of two prototypes belonging to the same class:

$$ML = \{((B_i(z), NB_i(z), ((B_j(z), NB_j(z) \in Z^{\,2})|\exists(l,m); l = m; (B_i(z), NB_i(z))\ \&\,\&\ (B_j(z), NB_j(z)\} \in Z^{\,l}\},$$

The cannon-link (CL) pairs can be built with the set of $(v.(v-1).p^2)$ pairs with two prototypes belonging to different classes.

$$CL = \{((B_i(z), NB_i(z)), (B_j(z), NB_j(z)) \in Z^{\,2})|\exists(l,m); l \neq m; (B_i(z), NB_i(z)) \in Z^{\,l}\ \&\ (B_j(z), NB_j(z)\} \in Z^{\,m}\}$$

4.1.3 Intuitionistic fuzzy Supervised Constraint score for selected feature subset

The target similarity matrix $\breve{\omega}^{sv}$ of Intuitionistic Fuzzy Supervised learning $\mathfrak{T}^{SV}(\mathcal{F}_k))$ is mathematically represented as

$$\breve{\omega}_{ij}^{sv} = \begin{cases} 1 & if\ (V_{i,}U_j) \in ML \\ 0 & if\ (V_{i,}U_j) \in ML \\ \omega_{ij}(\mathcal{F}_k) & else \end{cases}$$

$$V_{i,} = (B_i(z), NB_i(z) \qquad); \quad U_j = \ (B_j(z), NB_j(z))$$

In this proposed work to determine the similarity of any two records z_i, z_j is computed with the following Intuitionistic fuzzy similarity matrix is computed as

$$\omega_{ij} = \sqrt{\frac{1}{2}\sum_{i,j=1}^{n}(B_i(z)-B_j(z))^2 + (NB_i(z)-NB_j(z))^2 + (\varphi_i(z)-\varphi_j(z))^2}$$

$$\varphi_i(z) = 1-(B_i(z)+NB_i(z))$$
$$\varphi_j(z) = 1-(B_j(z)+NB_j(z))$$

In this proposed work the hesitation index $\varphi_i(z), \varphi_j(z)$ plays a vital role in handling uncertainty during the similarity measure among vague instances. With this for all the subset of features of $\omega_{ij}(\mathcal{F}_k)$ the weight matrix is generated.

$$\Delta_i = (B_i(z), NB_i(z), \varphi_i(z),)\ \Delta_j = (B_j(z), NB_j(z), \varphi_j(z))$$

$$\omega_{ij}(\mathcal{F}_k) = \exp\,(-\frac{\tau^2 {\Delta_i}^k, {\Delta_j}^k}{2\beta^2})$$

Where τ^2 refers to the intuitionistic fuzzy normalized similarity measure and β signifies scaling parameter. At last, the intuitionistic fuzzy constraint based on supervised learning $\mathfrak{T}^{SV}(\mathcal{F}_k)$) selects the potential subset features without considering the unlabelled datasets. The relevant subsets exhibit the characteristic such that $\omega_{ij}(\mathcal{F}_k)$ must satisfy the criteria the similarity score of two must-link records must be nearer to 1 and a cannot link records must be nearer to 0.

3.1.4 Intuitionistic Fuzzy Semi-Supervised Constrain score for feature selection

In Intuitionistic Fuzzy Semi-supervised learning, the constrain score is evaluated for both labelled and unlabelled chronic disease datasets. With the labeled datasets new must-link constraints are defined using the prototypes subsets and using unlabelled datasets are used for computing binary target similarity matrix $\breve{\omega}^{ssv}$

$$\breve{\omega}_{ij}^{ssv} = \begin{cases} 1 & if\ (V_{i,}U_j) \in ML^{SSV} \\ 0 & else \end{cases}$$

$$ML^{SSV} = \{(V_{i,}U_j) \in Z^2 |\ \exists\ l = 1,..,r\ (B_i(z), NB_i(z))\ \&\,\&\ (B_j(z), NB_j(z)\} \in Z^{\,l}\}$$

For clustering the unlabelled dataset, in this proposed work intuitionistic fuzzy K-NN is used for selecting the nearest neighbors which exhibit the same prototypes. And based on it the target similarity matrix is constructed.

4.2 Validation Using Deep Neural Network Classifier

In this proposed work, the feature subsets selected by the three different feature selection algorithms are validated using Deep Neural Network (DNN) [20] classifier. Multiple intermediate layers compute the large neural network using input and transform to output. The selected features are passed as input, the intermediate layers perform computation with weight and bias values, and these values get changed from layer to layer. The observed output is compared with the expected output and their error rate is passed to the learning model, back propagation is used to adjust the learning rate, weight, and bias values. The gradient descent method is used for the assignment of parameter values. This iterative process is accomplished on a trial-and-error basis until it meets the termination criteria or there are no other changes in the output.

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

The evaluation part analysis in detail the performance of the proposed Intuitionistic Fuzzy Constrained Feature Selection Algorithm (IFCFSA) on the detection of four different chronic disease dataset feature selections. In this work, four different chronic diseases namely Cancer [21], Diabetes [22], heart disease [23] and Stroke datasets [24] are used in this work for discovering the feature subset selection collected from Kaggle repository. The dataset undergoes five-fold cross validation, with the ratio of 80% instances as training dataset and 20% as testing dataset. The proposed model is deployed using Python software. The dataset with missing values is processed using the mean value imputation and the features with varying size is normalized using min-max normalization so that the all the values are converted to same range.

Table 1: Description of Chronic Diseases Dataset

Dataset	**No. of instances**	**No. of features**	**No. of classes**
Heart Disease	296	13	5
Stroke	510	12	2
Breast Cancer	699	9	2
Diabetes	768	8	2

Table 2: Deep Neural Network-based prediction of heart disease with Complete dataset and Dimensionality Reduced Feature Subsets

Method	**Accuracy**	**Precision**	**Recall**	**Dimensionality**	**Selected Feature Subset (%)**
Complete	79.24	0.79	0.82	13	100.0
ImRMR	82.63	0.85	0.87	10	76.9
CSFS	89.59	0.88	0.89	8	61.5
IFZ-CS	93.52	0.95	0.96	5	38.5

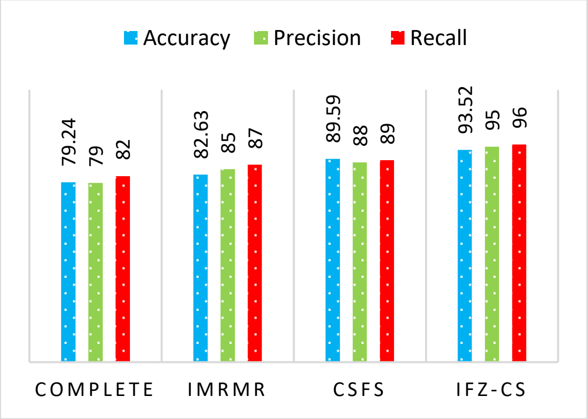

Figure 2 Comparison of feature selection models for heart disease prediction

Table 2 and Figure 2 illustrates the performance comparison of deep neural network with three different feature selection algorithms along with complete feature for predicting heart disease. The results show that the proposed intuitionistic fuzzy constrained based feature selection produced highest accuracy of 93.52% with the least dimensionality reduction of 5 feature subset. The proposed model focuses on hesitation index as important element in determining relevancy of each feature to the dependent class variable. The features with less redundancy and highly influential to prediction variable is effectively determined by their obtained intuitionistic fuzzy Constraint score value. Hence the proposed model performs better compared to other feature selection methods on heart diseases prediction.

Table 3: Deep Neural Network based prediction of Diabetes Disease with Complete dataset and Dimensionality Reduced Feature Subsets

Method	**Accuracy**	**Precision**	**Recall**	**Dimension**	**Feature Subset Selected (%)**
Complete	80.97	0.8	0.83	8	100.0
ImRMR	85.28	0.86	0.87	6	75.0
CSFS	88.69	0.89	0.9	5	62.5
IFZ-CS	94.16	0.96	0.97	3	37.5

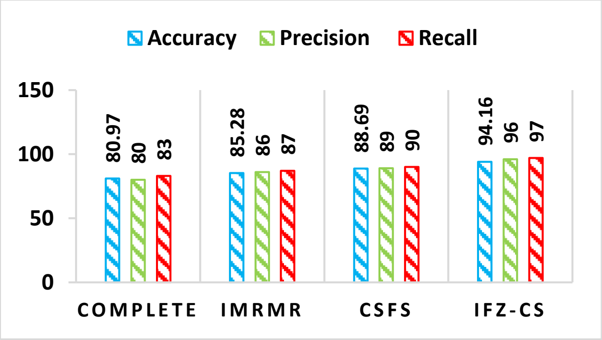

Figure 3 Comparison of feature selection models for diabetes disease prediction

The prediction results of the Deep Neural Network deployed on three different feature subsets and with the complete dataset, selected by three different feature selection methods for Diabetes prediction are shown in Table 3 and Figure 3. The complete dataset comprised 8 features, while using IFZ-CS it selects only 3 features to predict the diabetes presence. The result explores that the proposed model focuses on handling uncertainty in discovering the similarity among instances based on the prototypes determined by Intuitionistic fuzzy supervised and semi-supervised weight computation of pairwise features to improve the process of relevancy estimation more precisely. Thus, the proposed intuitionistic fuzzy constrained-based feature selection produced 94.16% accuracy, and while using a complete dataset and ImRMR due to redundancy and irrelevant features in chronic diseases dataset prediction they produce fewer results.

Table 4: Deep Neural Network-based prediction of Stroke Disease with Complete dataset and Dimensionality Reduced Feature Subsets

Method	**Accuracy**	**Precision**	**Recall**	**Dimension**	**Feature Subset Selected (%)**
Complete	83.51	0.84	0.82	12	100.0
ImRMR	85.67	0.87	0.89	10	83.3
CSFS	89.65	0.91	0.9	7	58.3
IFZ-CS	94.68	0.96	0.97	4	33.3

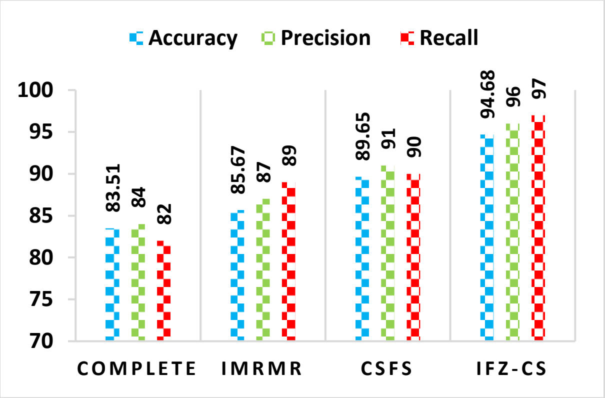

Figure 4: Comparison of feature selection models for stroke disease prediction

Table 4 and Figure 4 depict the result produced by the DNN classification model on three different feature subsets generated by ImRMR, CSFS, proposed IFZ-CS, and with a complete dataset of Stroke disease predictions. The potential features selected using the proposed IFCF achieve the highest rate of accuracy precision and recall. It also selects the least dimension of feature from the stroke dataset to predict the presence or absence of a stroke. The ability to represent each term in the form of degree of belongingness, non-belongingness, and hesitancy the problem of understanding the vagueness in similarity relationship among the feature sets both in supervised and semi-supervised are prominently determined by the proposed IFCF model. While other existing algorithms suffer from uncertainty conditions due to the redundant and irrelevant feature selection in chronic disease prediction.

Table 5: Deep Neural Network-based prediction of Cancer Disease with Complete dataset and Dimensionality Reduced Feature Subsets

Method	**Accuracy**	**Precision**	**Recall**	**Dimension**	**Feature Subset Selected (%)**
Complete	78.56	0.79	0.8	9	100.0
ImRMR	84.97	0.82	0.86	7	77.8
CSFS	86.95	0.87	0.89	6	66.7
IFZ-CS	93.82	0.95	0.96	5	41.7

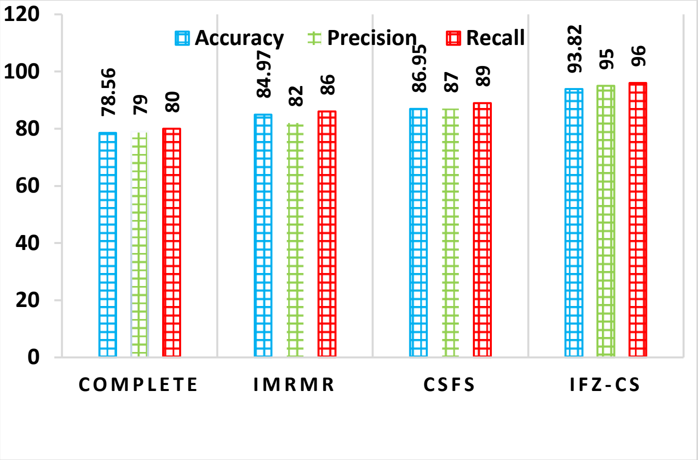

Figure 5 Comparison of feature selection models for cancer disease prediction

Table 5 and Figure 5 illustrate the performance of the proposed IFZ-CS achieves the highest rate of results in the prediction of cancer at its early stage, with the least subset of features. The deep neural network is used for the prediction of cancer and it used three different subsets of features generated by ImRMR, CSFS, and proposed IFZ-CS. This work also used the complete dataset to explore the consequences of predicting the process due to maximal redundancy and highly correlated independent variables resulting in the least accuracy rate. The supervised and semi-supervised intuitionistic fuzzy constraint-based similarity score enriches the process of selecting potential feature subsets and the deep neural network validates its selected features with the accuracy of 93.82% and it used the least feature subset of 41.7% to predict the cancer disease effectively.

6. CONCLUSION

The main motivation of this proposed work is to construct a robust feature selection model to overcome the primary issue of irrelevant, inconsistent, vague, and ambiguous feature reduction to improve the detection rate of chronic diseases. Hence, in this work, the intuitionistic fuzzy concept is fused with constraint score-based feature subset selection which works for both labeled and unlabelled datasets. The Intuitionistic fuzzy constraint score-based feature selection outperforms the existing constraint score similarity by focusing on the hesitation index of each attribute in terms of the degree of belongingness, non-belongingness, and hesitancy. The supervised and semi-supervised based IFZ-CS selects the feature subset by introducing the intuitionistic fuzzy target weight matrix to determine the strength of similarity among the attributes to achieve maximum relevancy and minimum redundancy in chronic disease detection. The performance evaluation is done on four different chronic disease datasets cancer, stroke, diabetes and heart disease and the validation of selected feature subset is done by the deep neural network classifier. The results proved that the complete dataset with its high correlation among the independent attributes affects the performance of the deep

***Article Received:** 05 July 2023 **Revised:** 25 August 2023 **Accepted:** 15 September 2023*

neural network. The existing minimal-redundancy-maximal-relevance (mRMR) fails to handle the uncertainty conditions which commonly occur in chronic disease datasets thus they achieved least results compared with IFZ-CS feature selection algorithm.

REFERENCES

1. M. U. Muhammad, R. Jiadong, N. S. Muhammad, M. Hussain, I. Muhammad, Principal Component Analysis of categorized polytomous variable-based classification of diabetes and other chronic diseases, Int. J. Environ. Res. Public Health, vol. 16, no. 19, pp. 3593, 2019.
2. Jadhav S, He H, Jenkins K, Information gain directed genetic algorithm wrapper feature selection for credit rating. Appl Soft Comput 69:541–553, (2018).
3. B. A. Tama, S. Im , S. Lee, "Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble", *BioMed Res. Int.*, vol. 2020, pp. 1-10, 2020
4. Nitin Chopde, Rohit Miri, Effects of Features Selection and Classification on Various Chronic Diseases Detection, International Journal of Advanced Science and Technology, 29(06), 6255 – 6261, 2020
5. EL-Rahman, S.A., Saleh Alluhaidan, A., AlRashed, R.A, Chronic diseases monitoring and diagnosis system based on features selection and machine learning predictive models. *Soft Comput* **26**, 6175–6199 (2022).
6. Junaid Rashid, Saba Batool, Jungeun Kim, Muhammad Wasif Nisar, Amir Hussain, Sapna Juneja, Riti Kushwaha, Front. Public Health, Digital Public Health, Volume 10 – 2022
7. J. Umamageswaran, G. Elangovan, A. V. Kalpana, G. Indumathi; Chronic kidney disease prediction with feature selection and extraction using machine learning. AIP Conference Proceedings 3 October 2022; 2519 (1): 030092
8. S. J. Sushma, Tsehay Admassu Assegie, D. C. Vinutha, S. Padmashree, An improved feature selection approach for chronic heart disease detection, Bulletin of Electrical Engineering and Informatics, Vol. 10, No. 6, December 2021, pp. 3501~3506
9. Mishra S, Mallick P.K.; Tripathy, H.K.; Bhoi, A.K.; González-Briones, A. Performance Evaluation of a Proposed Machine Learning Model for Chronic Disease Datasets Using an Integrated Attribute Evaluator and an Improved Decision Tree Classifier. *Appl. Sci.* 2020, *10*, 8137.
10. Divya Jain, Vijendra Singh, A two-phase hybrid approach using feature selection and Adaptive SVM for chronic disease classification, International Journal of Computers and Applications, 43:6, 524-536, 2021.
11. Hegde S., Mundada M.R., Early prediction of chronic disease using an efficient machine learning algorithm through adaptive probabilistic divergence-based feature selection approach, International Journal of Pervasive Computing and Communications, Vol. 17 No. 1, pp. 20-36, 2021.
12. Ebiaredoh-Mienye, S. A., Swart, T. G., Esenogho, E., Mienye, I. D. A Machine Learning Method with Filter-Based Feature Selection for Improved Prediction of Chronic Kidney Disease, Bioengineering, vol. 9,8 350, 2022
13. Esenogho E., Mienye I.D., Swart T.G., Aruleba K., Obaido G. A Neural Network Ensemble with Feature Engineering for Improved Credit Card Fraud Detection, 2022;10:16400–16407.
14. Saranya K R, Feature Selection and Classification Algorithms for Chronic Disease Prediction Using Machine Learning Techniques, International Journal of Science and Research, Volume 12 Issue 2,2023
15. Xie, S., Zhang, Y., Lv, D, . A new improved maximal relevance and minimal redundancy method based on feature subset. *J Supercomput* **79**, 3157–3180 (2023).
16. Abderezak Salmi, Kamal Hammouche , Ludovic Macaire, Similarity-based constraint score for feature selection, Knowledge-Based Systems 209 (2020) 106429
17. P. A.Ejegwa, S.O. Akowe, P.M. Otene, J.M. Ikyule, An Overview On Intuitionistic Fuzzy Sets, International Journal Of Scientific & Technology Research Volume 3, Issue 3, 2014
18. Dan, S.; Kar, M.B.; Majumder, S.; Roy, B.; Kar, S.; Pamucar, D, Intuitionistic Type-2 Fuzzy Set and Its Properties. Symmetry 2019, 11, 808.
19. K. Meena, Lija Ponnappen, An Application of Intuitionistic Fuzzy Sets in Choice of Discipline of Study, Global Journal of Pure and Applied Mathematics, Volume 14, Number 6 (2018), pp. 867–871
20. Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8, 53 (2021).
21. https://www.kaggle.com/rishidamarla/cancer-patients-data
22. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
23. https://www.kaggle.com/datasets/mathchi/diabetes-data-set
24. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset