Optimized Class Level Attraction Measure Based Machine Learning for Improved Disease Prediction Using Support Vector Machine

Dr. Nidhi Mishra[1], **Sushree Sasmita Dash**[2]

[1]*Associate Professor, Faculty of CS & IT, Kalinga University, Naya Raipur, Chhattisgarh, India, ku.nidhimishra@kalingauniversity.ac.in*
[2]*Assistant Professor, Faculty of CS & IT, Kalinga University, Naya Raipur, Chhattisgarh, India, ku.sushreesasmitadash@kalingauniversity.ac.in*

The problem of disease prediction has been approached with several techniques in literature. The methods perform disease prediction according to set of symptoms and different methods like decision tree, Bayesian classification, Artificial neural network are used in this problem. However, the methods suffer to achieve higher performance in disease prediction. Towards this issue, an efficient Optimized Class Level Attraction Measure based disease prediction model with SVM (OCLAM-SVM) is presented in this article. The method preprocesses the data set and extracts the features to train the support vector machine. The support vector machine is designed to measure the support value towards various disease classes by computing OCLAM value. Based on the OCLAM value, the method identifies the class of disease to perform disease prediction.

Keywords: Machine Learning, Disease prediction, SVM, OCLAM-SVM.

1. Introduction

The modern society struggle with number of diseases in their life. Some of the diseases make their life to end and some of them produce temporary illness. Any disease has been identified based on the set of symptoms. For example, the presence of diabetes is predicted based on the blood sugar, BMI and HbA1C. By considering the above mentioned symptoms, the presence of diabetes can be predicted. Similarly, any disease can be predicted with set of symptoms. Predicting the disease at the early stage support the life management in a most efficient way.

The disease prediction is the process of predicting the possibility of disease according to the symptoms. However the performance of the disease prediction algorithm is greatly depend on the set of features considered and the volume of data being used. In order to achieve higher disease prediction, the method should consider maximum features and large volume of data samples. The data mining algorithms are used in this case like ensemble learning. The method generates set of ensembles from the data point and computes ensemble similarity. However, the performance of the methods is not convincing one.

The machine learning algorithms are used in different scientific problem. There are number of methods available in literature like K-means, support vector machine, decision tree, Bayesian classification and particle swarm optimization. Each method has their own merit in disease prediction. However, the machine learning algorithms are capable of handling huge volume of data which support the improvement of disease prediction.

By considering all these, an efficient Optimized Class level Attraction Measure based SVM model is presented towards disease prediction. The model estimates class level attraction measure according to different features and symptoms to perform disease prediction. The model trains the support vector machine with number of records and estimates OCLAM measure to perform disease prediction.

2. Related Works:

There exist number of approaches towards disease prediction and this section details set of methods around the problem.

A deepwalk and random forest based Metabolite-disease association prediction model is presented in [1], where the DWRF approach measure the semantic and information similarity to perform disease prediction.

A Higher-Order Proximity-Based MiRNA-Disease Associations Prediction (HOP_MDA) is presented in [2], which predict the association among miRNA and the diseases according to the interaction. A local random walk based human LncRNA disease prediction model is presented in [3], which infer potential association among human IncRNA and disease. The same IncRNA disease association prediction is presented in [4], which integrate the semantics of multiple meta paths of pairwise attributes with graph to perform disease prediction. A machine learning based prediction model is presented in [5], which uses fuzzy GDBT algorithm with fuzzy logic and gradient boosting decision tree to perform disease prediction. An automatic disease prediction and crop selection model is presented in [6], for the support of smart agriculture which uses CNN model to analyze the disease prediction. A Alzheimer disease progression prediction model is presented in [7], which combines 3D multi-information generative adversarial network (mi-GAN) to predict the disease. A

Graph Triple-Attention Network based disease prediction model is presented in [8], which predict the association propensities between lncRNAs and diseases. A smart watch based machine learning framework is presented in [9], which use different machine learning algorithm to perform disease prediction. A Webserver Integrating Database and Prediction Tools for Microbe-Disease Associations (MDADP) is presented in [10]. An Alzheimer disease prediction model is presented in [11], which works based on the population and brain age estimation towards prediction. A soil sensor based plant disease prediction model is presented in [12]. An arterial disease prediction on inflammatory bowel disease is presented in [13], which use the arterial event and features to perform disease prediction. A pair wise attribute and neighbor topological structure based prediction model is presented in [14], which learn and integrate pairwise information to measure the similarity among the drugs to perform disease prediction.

3. Optimized Class Level Attraction Measure based disease prediction model with SVM (OCLAM-SVM):

The proposed OCLAM-SVM model reads the medical data set and applies preprocessing technique to normalize the data set. Further the method extracts the features of the records to generate feature vector. Generate feature vector set has been trained with support vector machine. At the training phase, the features of the test sample has been extracted and passed to the support vector machine. The support vector machine estimates optimized class level attraction measure towards various classes of diseases. Based on the OCLAM value, the method identifies the class of disease as result. The detailed working of the proposed model is explained in this section.

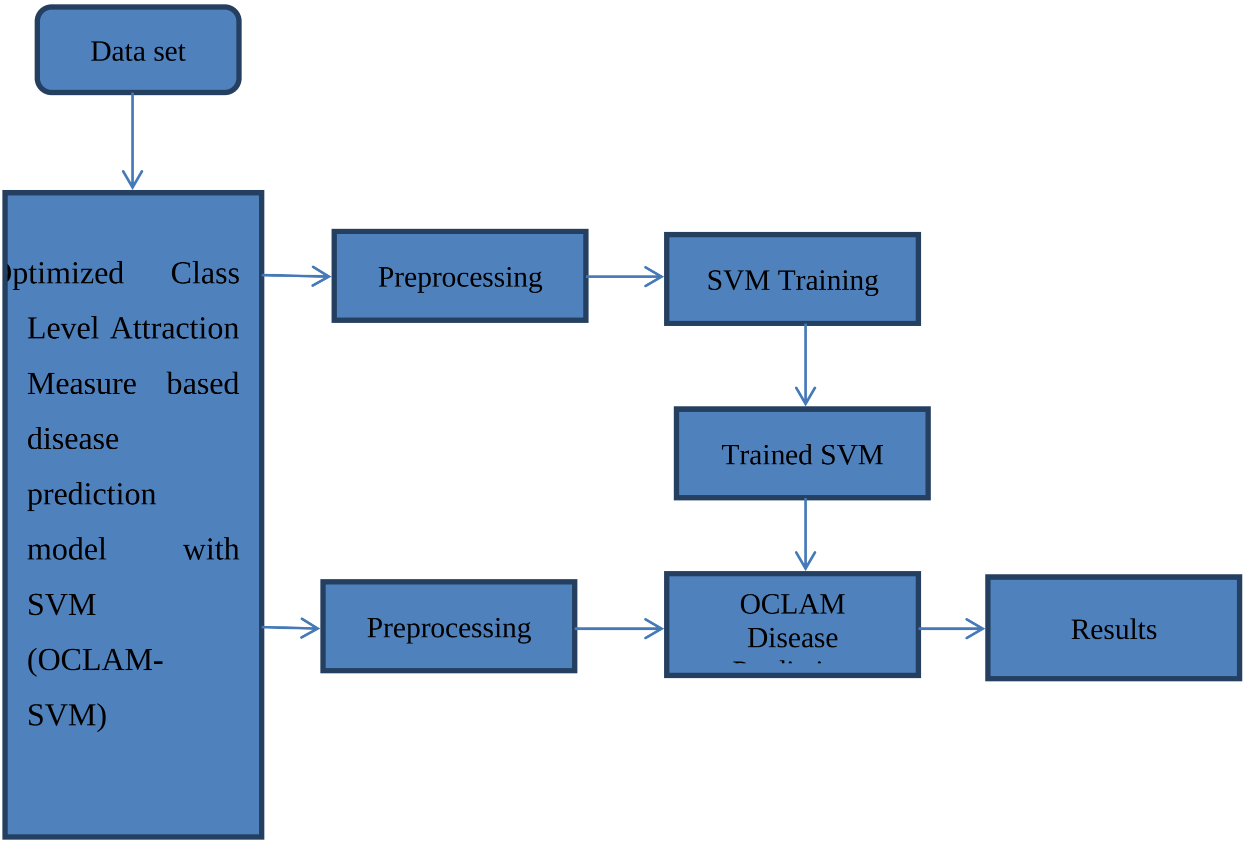

Figure 1: Architecture of Proposed OCLAM-SVM Model

The working procedure of OCLAM-SVM model is sketched in Figure 1, and briefed in detail in this section.

Preprocessing:

The medical data set given has been preprocesses to remove the noisy features and records from the set. Initially, the method identifies the set of features from the records of the data set. According to the features identified, the method checks the records for the presence of all the features and values. If anything found missing then it has been removed from the set. The normalized data set has been used towards disease prediction.

Algorithm:

Given: Medical Data set mes

Obtain: Preprocessed set Prs, feature vector set Fvs

```
Start
    Read Mes.
    Find feature set Fes = $\underset{i=1}{\overset{size(Mes)}{Fes \cup (Features\ (Mes(i)) \ni Fes)}}$
    For each sample s
        If $s \in \forall Features(Fes))$ then
            Prs = Prs ∪ $s$
            Generate Feature Vector Fv = $\sum FeatureValues \in s$
            Fvs = Fvs ∪ $Fv$
        Else
            Mes = Mes ∩ $s$

        End
    End
Stop
```

The preprocessing algorithm removes the noisy records from the set and generates feature vector set to support disease prediction.

SVM Training

In this phase, the support vector machine maintains two binary classes. With the preprocessed data set and the feature vector set, the method computes Class Level Disease Rate (CLDR) value towards various class of disease. According to the value of CLDR value, the method index the feature vector to the class identified. The trained model has been used to perform disease prediction.

```
Algorithm:
Given: Preprocessed set Prs, Feature vector set Fvs
Obtain: SVM
Start
    Read Prs and Fvs.
    Initialize SVM.
    For each class of SVM
        Index set of random vectors.
    End
    For each feature vector fv
        For each class c
            For each feature vector fv1
                Compute Disease score DS = $\frac{\overset{size(fv)}{\underset{i=1}{Count(Dist(Fv(i),Fv1(i))<Th)}}}{size(Fv)}$
            End
            Compute class level disease rate CLDR = $\frac{\sum DS}{size(c)}$
        End
        Class c = choose the class with maximum CLDR.
        Index the sample to the class identified.
    End
Stop
```

The SVM training algorithm computes the class level disease rate to identify the class of sample and indexed. Trained machine has been used to perform disease prediction.

OCLAM Disease Prediction:

The test sample is preprocessed and its features are extracted to produce the feature vector. The feature vector generate has been passed to support vector machine. The SVM computes optimized class level attraction measure (OCLAM) is measured towards various classes. Based on the OCLAM, the method identifies the class of disease as result.

```
Algorithm:
Given: SVM, test sample Ts
Obtain: Disease class Dc
```

```
Start
    Read SVM and Ts.
    [Ps, Fv] = preprocessing (Ts)
    For each class c
        Compute OCLAM = $\frac{\sum_{i=1}^{size(c)} \frac{\sum_{j=1}^{size(Fv)} Fv(j)==C(i)(j)}{size(Fv)}}{Size(c)}$
    End
    Dc = Choose the class with maximum OCLAM.
Stop
```

The disease prediction algorithm computes the OCLAM measure for various disease class to identify the possible disease.

4. Results and Discussion:

The proposed OCLAM-SVM disease prediction algorithm has been implemented using Matlab. The performance of the model is evaluated using Kaggle diabetes data set and obtained results are compared with the result of others.

Table 1: Evaluation Details

Key	Value
Tool Used	Matlab
Data set	Kaggle
Features	32
No of samples	735

The evaluation details used to perform disease prediction is presented in Table 1, and the methods are analyzed for their performance in various metrics.

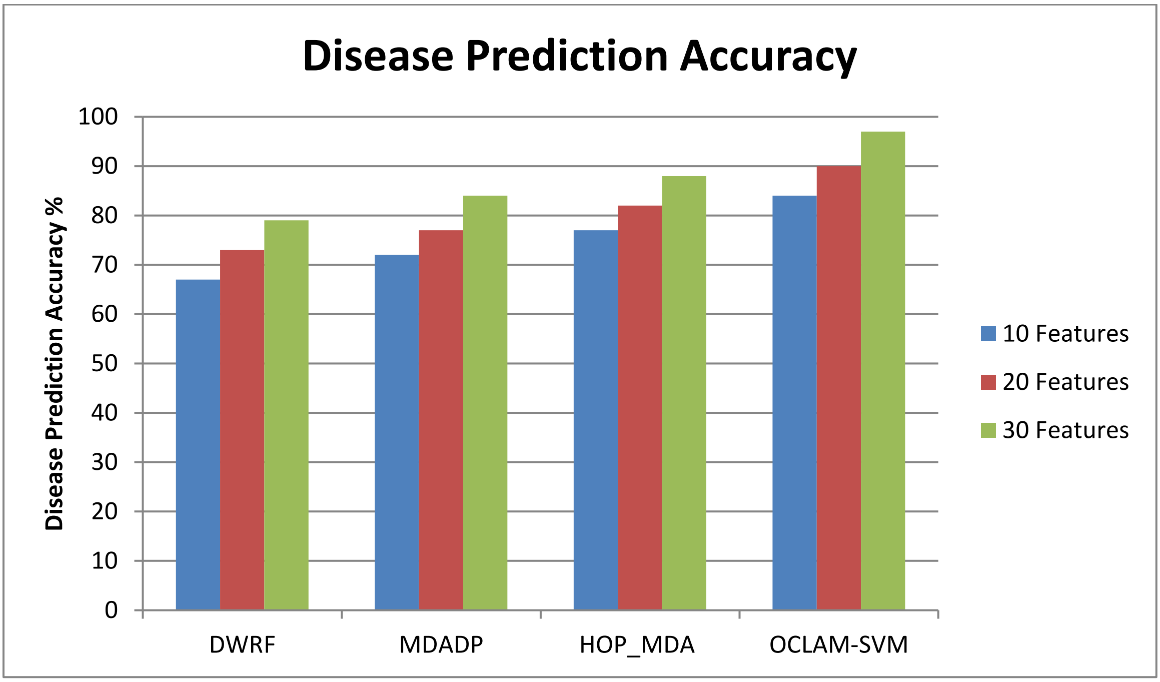

Figure 2: Disease Prediction Accuracy

The accuracy in disease prediction is measured and compared in Figure 2, where the OCLAM-SVM model introduces higher disease prediction accuracy than others.

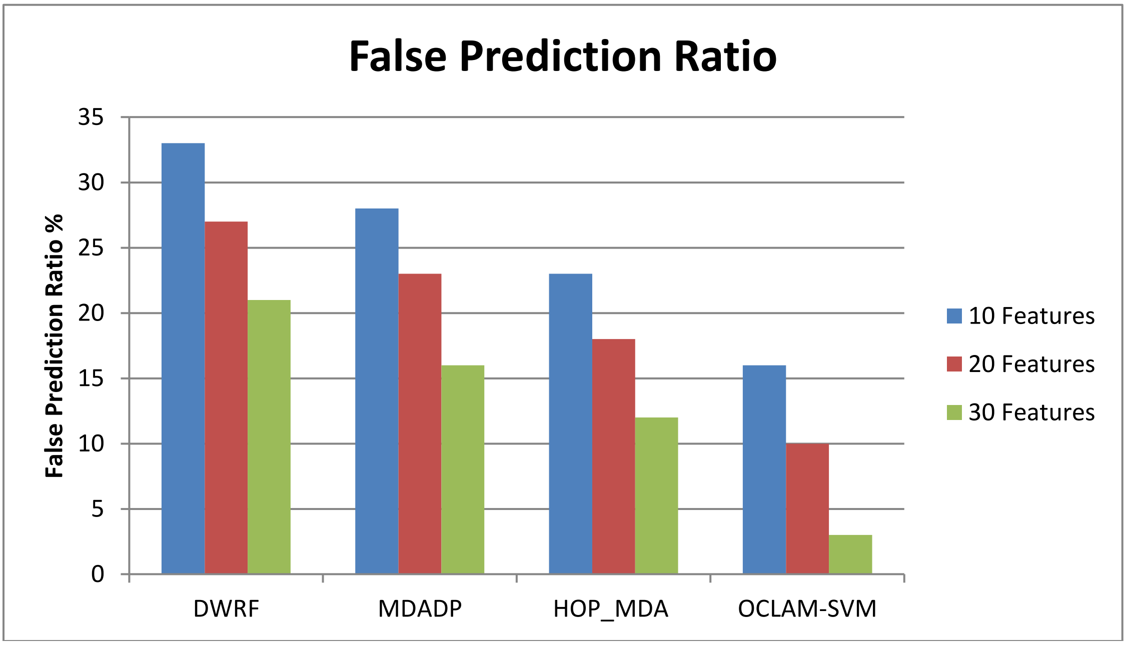

Figure 3: False Prediction Ratio

The ratio of false prediction has been measured and presented in Figure 3, where OCLAM-SVM model introduces less false ratio than others.

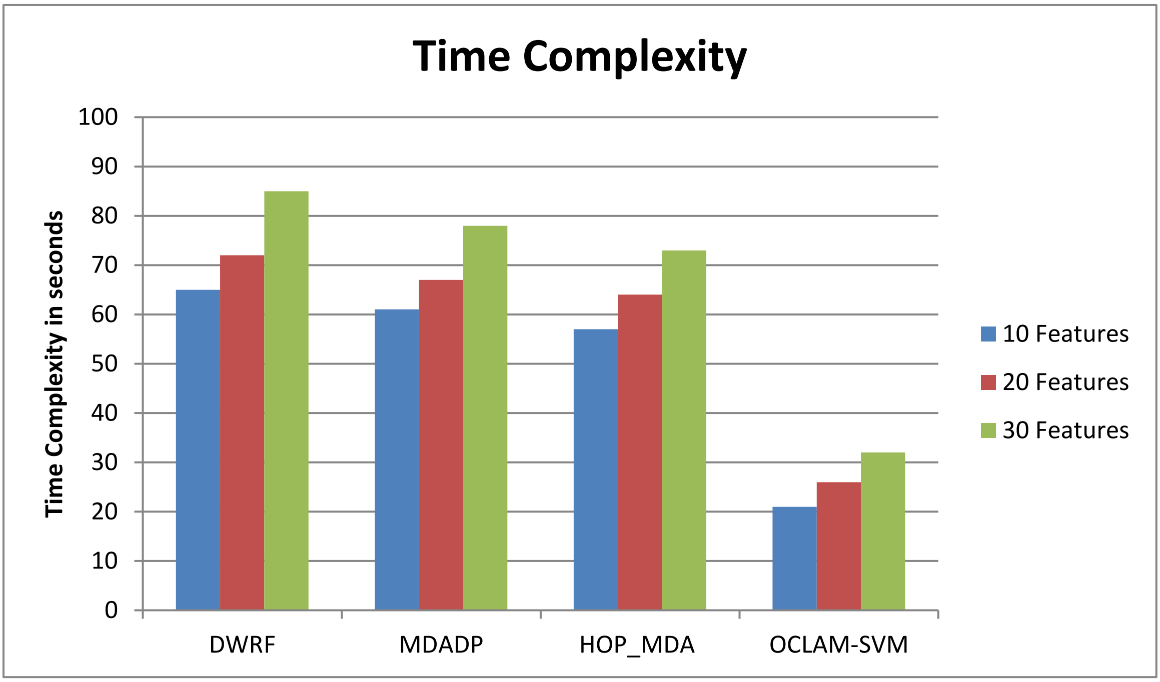

Figure 4: Time complexity

The time complexity incur on disease prediction is measured and compared in Figure 4, where OCLAM-SVM model introduces less time complexity compare to other methods.

5. Conclusion:

This paper presented a novel Optimized class level attraction measure with SVM based disease prediction model. The model reads the medical data set and applies preprocessing technique to normalize the data set. Further the method extracts the features of the records to generate feature vector. Generate feature vector set has been trained with support vector machine. At the training phase, the features of the test sample has been extracted and passed to the support vector machine. The support vector machine estimates optimized class level attraction measure towards various classes of diseases. Based on the OCLAM value, the method identifies the class of disease as result. The method introduces higher performance in disease prediction with less time complexity.

References

1. J. Tie, X. Lei and Y. Pan, "Metabolite-disease association prediction algorithm combining DeepWalk and random forest," IEEE (TS&T), Volume. 27, Number 1, pp. 58-67, 2022.
2. S. Zhang, J. Li, W. Zhou, T. Li, Y. Zhang and J. Wang, "Higher-Order Proximity-Based

MiRNA-Disease Associations Prediction," IEEE/ACM (TCB&B), Volume. 19, Number 1, pp. 501-512, 2022.
3. J. Li et al., "A Novel Approach for Potential Human LncRNA-Disease Association Prediction Based on Local Random Walk," IEEE/ACM (TCB&B), Volume. 18, Number 3, pp. 1049-1059, 2021.
4. P. Xuan et al., "Semantic Meta-Path Enhanced Global and Local Topology Learning for lncRNA-Disease Association Prediction," IEEE/ACM (TCB&B), Volume. 20, Number. 2, pp. 1480-1491, 2023.
5. X. Yuan, J. Chen, K. Zhang, Y. Wu and T. Yang, "A Stable AI-Based Binary and Multiple Class Heart Disease Prediction Model for IoMT," IEEE (TII), Volume. 18, Number 3, pp. 2032-2040, 2022.
6. V. Udutalapally, S. P. Mohanty, V. Pallagani and V. Khandelwal, "sCrop: A Novel Device for Sustainable Automatic Disease Prediction, Crop Selection, and Irrigation in Internet-of-Agro-Things for Smart Agriculture," IEEE (SJ), Volume. 21, Number 16, pp. 17525-17538, 2021.
7. Y. Zhao, B. Ma, P. Jiang, D. Zeng, X. Wang and S. Li, "Prediction of Alzheimer's Disease Progression with Multi-Information Generative Adversarial Network," IEEE (JB& HI), volume. 25, number 3, pp. 711-719, 2021.
8. P. Xuan, L. Zhan, H. Cui, T. Zhang, T. Nakaguchi and W. Zhang, "Graph Triple-Attention Network for Disease-Related LncRNA Prediction," IEEE (JB&HI), Volume. 26, Number 6, pp. 2839-2849, 2022.
9. S. T. Himi, N. T. Monalisa, M. Whaiduzzaman, A. Barros and M. S. Uddin, "MedAi: A Smartwatch-Based Application Framework for the Prediction of Common Diseases Using Machine Learning," in IEEE Access, Volume. 11, pp. 12342-12359, 2023.
10. L. Wang et al., "MDADP: A Webserver Integrating Database and Prediction Tools for Microbe-Disease Associations," IEEE (JB&HI), Volume. 26, Number 7, pp. 3427-3434, 2022.
11. W. Liu, Q. Dong, S. Sun, J. Shen, K. Qian and B. Hu, "Risk Prediction of Alzheimer’s Disease Conversion in Mild Cognitive Impaired Population Based on Brain Age Estimation," IEEE (TNS&RE), Volume 31, pp. 2468-2476, 2023.
12. M. Kumar, A. Kumar and V. S. Palaparthy, "Soil Sensors-Based Prediction System for Plant Diseases Using Exploratory Data Analysis and Machine Learning," IEEE (SJ), Volume. 21, Number 16, pp. 17455-17468, 2021.
13. D. Chicco and G. Jurman, "Arterial Disease Computational Prediction and Health Record Feature Ranking Among Patients Diagnosed With Inflammatory Bowel Disease," in IEEE Access, Volume. 9, pp. 78648-78657, 2021.
14. Y. Song, H. Cui, T. Zhang, T. Yang, X. Li and P. Xuan, "Prediction of Drug-Related Diseases Through Integrating Pairwise Attributes and Neighbor Topological Structures," IEEE/ACM (TCB&B), Volume. 19, number 5, pp. 2963-2974, 2022.