Modelling An Effectual Feature Selection Approach For Predicting Down Syndrome Using Machine Learning Approaches

[1]Dr.M.Jaganathan, [2]Mr.R.Gopal, [3]V.R.Kiruthika,

[1]Assistant Professor-II, *Computer Science and Engineering, Bannari Amman Institute of Technology, Erode, Tamil Nadu*
[2]Assistant Professor, *Computer Science and Engineering, Bannari Amman Institute of Technology, Erode, Tamil Nadu*
[3]Assistant Professor *Computer Science and Engineering, Bannari Amman Institute of Technology, Erode, Tamil Nadu*

Abstract- Down Syndrome (DS) is a genetic disorder which is caused due to the occurrence of third copy of chromosome 21. For DS, pre-natal screening is the primary component and it is suggested to be offered universally to women irrespective of their background and age. Machine Learning (ML) plays an essential role in predicting the severity of the disease in earlier stage with the features related to DS. It gains a considerable attention in performing predictive analysis for various medical applications. Therefore, the effectual and appropriate diagnosis of DS is a significant challenge for medical practitioners and experts. The ultimate target of this work is to initiate an accurate and non-invasive diagnostic process for predicting DS and to reduce the cost of basic prenatal diagnosis. An effectual ML approach is developed in this work to diagnose DS. Here, L1-norm based Support Vector Regression (L1-SVR) for feature selection is applied for selecting the highly related and appropriate features for accurate classification of DS from normal people. The proposed L1-SVR generates a newer feature subset from the available dataset based on its feature weighted value. The performance metrics like sensitivity, specificity, accuracy, F1 score, precision are evaluated for evaluation. The optimal accuracy attained with this finest subset of chosen features is due to diverse contributions of the DS features. The experimental outcomes of this study recommend that the anticipated model is applied for appropriate prediction of DS and can be applied for making proper decision during the critical condition. Recently, computer aided decision support system plays a significant role in assisting DS prediction. The proposed L1-norm SVM pretends to fulfill the gap among the feature selection process and classification using the available data by properly fulfilling the experimental design. The simulation is done with MATLAB simulation environment.

Keywords- Down syndrome, machine learning, Support Vector Regression, feature selection, optimal subset features

1. INTRODUCTION

Down syndrome is a common genetic chromosomal disorder that is characterized as Hsa21 trisomy, this result in errors on Robertsonian or non-dysjunction translocation among 21 chromosomes and other chromosomes [1]. With the report given by various investigators, DS affects one out of 700 live birth cases all over the world. The individuals/subjects who are affected by DS shows alternations in their body with dysmorphic facial features, variations in co-occurring medical conditions, ID [2]. It also includes leukemia, congenital heart defects, and Alzheimer's disease. Due to the frequency of occurrence and severity, the prediction and diagnosis of DS is concentrated by various researchers. Based on the Hsa 21 trisomy and genes of other chromosomes affects DS as in Fig 1, protein expression dosage that is provided based on the subset of the encoded genes which are increasing. The expression of this protein includes RNA splicing factors, adhesion molecules, cell surface receptors, protein modifiers, transcription factors and other related components of various bio-chemical pathways that cause Learning and Memory deficiencies [3].There are huge investigations have been performed mouse to analyze the severity of the human diseases. It plays a pre-dominant role in prediction and diagnosis of DS for further treatment [4]. Moreover, it is extremely complex to design a prediction approach for DS for other organisms as orthologs of Hsa21 genes can be only mapped with the chromosomes of mouse, i.e., 10, 16 and 17 respectively [5]. The chromosome of mice consists of 88 orthologs with Has 21 protein coding genes and microRNA genes [6]. At present, it is some popular in pre-clinical evaluations of treatments for DS [7]. This work concentrates on analyzing the features on the DS using Machine Learning approaches. Recently, ML plays pre-dominant role in disease prediction. Here, the subset features are chosen from protein expression data that comprises of various classes alike of mice, i.e., 77 protein expression levels, cortex control, and Ts65 Dntrisomic. For feature selection process, L1-norm based Support Vector Regression (SVR) is applied and proteins are chosen based on the enhancements of model. After selecting appropriate features, it is fed as an input to the classifier model. However, this work specifically concentrates on feature selection alone as it leads to degradation of classifier performance when not chosen properly and lacks in accuracy rate [8]. The proposed L1-SVR is shows better generalization and separates the class labels effectually by reducing the classification risks. The performance of L1-SVR is compared with prevailing approaches like k-NN, SVM, Adam, SGDM, and RMSPROP respectively. The evaluation metrics gives higher performance than other models. In this work, the effectual subset feature selections are done to acquire appropriate classification data models than the subset chosen from the previous studies. The chosen subset features are used to examine the influence of learning and memory deficiencies and it can be used for effectual classification and prediction process of DS in earlier stage and to treat it accordingly.

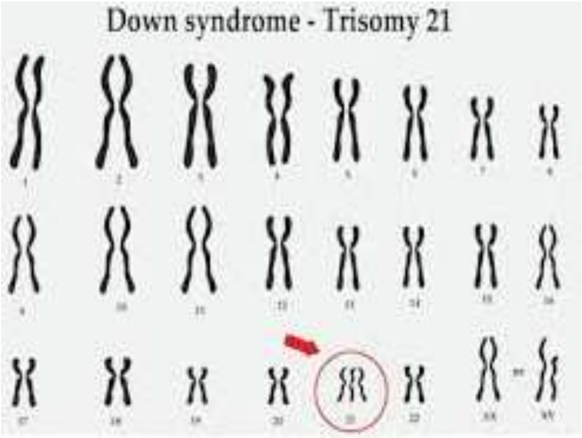

Fig 1: Trisomy 21 for DS

The work is partitioned as: section 2 is background investigations associated with DS and use of ML algorithms for prediction; section 3 is L1-norm based SVR demonstration for effectual feature selection; section 4 is numerical results with comparison of other prevailing models; section 5 is conclusion with future research directions.

2. RELATED WORKS

The DS prediction is performed during the pregnancy condition or after the child birth. Shelhamer et al., in [9], performed DS screening which is suggested universally to protect the pregnant woman ad it is considered as a critical element during the ante-natal care. After the child birth, DS is predicted with the occurrence of various typical facial features. Some of these features comprises of flat nasal bridge, upslanted palpebral fissures, wider eye space, small nose and ears and protruding tongue. The chromosome test termed as non-classical Karyotype test is carried out to confirm the occurrence of DS. Moreover, the prediction of non-classical occurrence of DA is limited by clinical experts based on their experiences. These chromosome tests are computationally complex, expensive and time-consuming and various remote health institutions posses no proper access towards the technologies. Hence, the adoption of computerized systems between the health professionals is considered to be increasingly essential.

The recent advancements in computer vision and machine learning approaches give opportunities for development in various medical fields as stated by Yi et al., [10]. The task performance like localization, object detection, recognition, and segmentation process relies over the available public datasets with drastic enhancements for past few years. Some essential types of learning approaches are neural network model which deals with various layers with enormous amount of trainable parameters which is updated/revised constantly using back-propagation algorithm to reduce the loss among the targets and outputs during training process. In medical field, ML shows significant improvements in lesion segmentation and disease prediction based on its powerful ability for extraction of features. The distinguishing facial features rely over DS might give an opportunity for automatic prediction as described by Hassner et al., in [11]. Recently, some investigations have been performed to predict the DS cases with 3D or 2D facial images.

Eidinger et al., in [12], proposed a model with the use of texture bio-markers and facial geometrics for DS prediction using 2D facial images. The facial features are represented with certain geometric features that rely over facial anatomical landmarks, local binary patterns, and local texture feature based contourlet transform. The abnormal and normal cases are differentiated using ML approaches that include k-NN and SVM. Moreover, this method requires manually extracted geometric features from various patients and dataset which includes 24 DS cases and 24 normal cases. Based on this analysis, there is no proper publication which is associated with fully automatic prediction of DS with facial recognition technologies specifically known as NN as depicted by David et al., in [13].

Currently, ML approaches have acquired a most pre-dominant attention and have shown a wider range of prediction with cancer diagnosis and other common disease prediction. It is also used for understanding the complexity of disease severity and to produce disease based medication from various available clinical and literature data repository. Moreover, some applications of ML over DS screening process is reported with highly imbalanced feature correlated data. Tara et al., in [14], explains about the ANN with under-sampling strategy (under-sampled ANN) with available dataset which is offered by Fetal Medicine Foundation to identify chromosomal abnormalities. Moreover, the provided model is promoted globally as there are huge differences among the serum markers concentrations which are associated with various races and regions. While considering DS incidence, data collection from 14/10,000 live births for generation of hungry data ANN model with extensive analysis. Similarly, lightweight ML approaches are trained with lesser number of DS samples which is highly suggested and recommended for modeling a special design for certain regional and ethnic available datasets [16].

1. *Methodology*

This section concentrates on modeling of L1-norm SVR for feature selection. Initially, the dataset is considered from MSS online available dataset attained from Center of Prenatal diagnosis. It comprises of 108 positive and 100,244 negative cases those outcomes in imbalanced data ratio of 1:928 [15]. The anticipated L1-norm SVR is used for handling the imbalanced data effectually. The flow diagram of the L1-norm SVR is given in Fig 2.

a) Dataset description

The provided dataset is specified by 22 feature vectors. These features show higher significance as it is considered during the pregnancy state. For example, ultra-sonographic markers, unconjugated estriol, $\beta -$ human chorionic gonadotropin, biochemical pregnancy based plasma protein, bipartial diactoc. These features are associated with various physiological and historical data of pregnant women. These data include weight maternal age, nationality, menstrual cycle, abnormal pregnancy history and so on. The biochemical markers are normalized with various medians with effectual normalization techniques. Some features like fetal chromosome karyotype and telephone follow-up are essential based on real labels of samples.

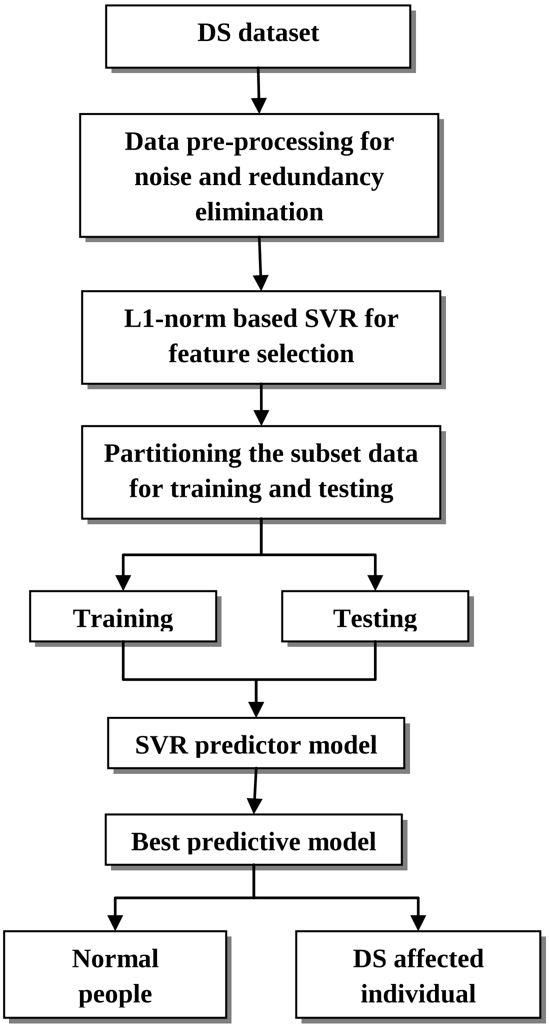
DS dataset
Data pre-processing for noise and redundancy elimination
L1-norm based SVR for feature selection
Partitioning the subset data for training and testing
Training
Testing
SVR predictor model
Best predictive model
Normal people
DS affected individual

Fig 2: Flow diagram of proposed model

b)Support Vector Regression (SVR)

In general, Support Vector Machine (SVM) is a popular learning algorithm which is used for both regression and classification. There is a significant model to perform regression termed as $\ell_1 - norm$ SVR which has the ability to predict the most critical features to perform regression [17]. This model is extremely essential for various medical applications where huge noisy and redundant features are identifies, i.e. regression applied for biological data. Generally, $\ell_1 - norm$ is considered as a linear programming problem. However, some standard programming solvers like simplex algorithms are used for resolving it. Moreover, in large-scale problems, certain effectual algorithms are accessible and popular for handling regression problem.

C)$\boldsymbol{\ell_1 - norm}$ based SVR for DS prediction

Consider, an unknown system $f: R^{n-1} \rightarrow R$ which transforms the given input data vectors of DS dataset to real number $f(\vec{x})$. The target of $\ell_1 - norm$ based SVR is to evaluate $f(\vec{x})$ by considering the $'m'$ training samples $(\vec{x_1}, f(\vec{x}_1)), (\vec{x_2}, f(\vec{x}_2)), \dots, (\vec{x_m}, f(\vec{x}_m))$. In linear $\ell_1 - norm$ SVR, $'f'$ is expressed as in Eq. (1):

$$f(\tilde{x}) = \tilde{w}^T \tilde{x} + b \quad (1)$$
$$\text{With } \tilde{w} \epsilon R^{n-1}; b \epsilon R$$

Here, $x = \begin{pmatrix} \tilde{x} \\ 1 \end{pmatrix} \epsilon R^n$ and $w = \begin{pmatrix} \tilde{w} \\ b \end{pmatrix} \epsilon R^n$ is expressed as homogeneous form as in Eq. (2):

$$f(x) = w^T x \quad (2)$$

The feature vector redundancy problem is handled by $\ell_1 - norm$ based SVR as in Eq. (3):

$$\min \|w\|_1 \quad (3)$$
$$\text{Where,}$$
$$y_i = w^T x_i; \qquad i = 1,2, \dots, m$$

For real-world medical applications, the provided output is corrupted using various unknown distributed noises. The loss function $L(x, y, f(x))$, and penalty term $e^T \xi$ is initiated to deal with the noise. This is expressed as in Eq. (4):

$$\min \|w\|_1 + C e^T \xi \quad (4)$$
$$L(x_i, y_i, f(x_i)) \leq \xi_i; i = 1,2, \dots, m; \ \xi_i >= 0$$

Here, $C > 0$ is a constant for handling the trade-off among the regularization and deviation tolerance. The loss function is suited for all problems. With this, insensitive loss function is expressed as in Eq. (5):

$$L(x, y, f(x)) = \max\{|y - f(x)| - \epsilon, 0\}$$

It is generally considered loss function where ϵ is pre-specified values. With this loss function, the formula is expressed as in Eq. (6):

$$\min \|w\|_1 + C e^T \xi \quad (6)$$
$$|y_i - w^T x_i| \leq \epsilon + \xi_i; i = 1,2, \dots, m; \ \xi_i. = 0$$

This expression is provided for handling linear programming problem and various effectual algorithms are also used for handling the distributed noise over the input. The Gaussian Loss Function is expressed as in Eq. (7):

$$L(x, y, f(x)) = \frac{1}{2} (y - f(x))^2 \quad (7)$$
$$\min \|w\|_1 + C e^T \xi \quad (8)$$
$$(y_i - w^T x_i)^2 \leq \xi_i; \quad i = 1,2..m$$

These kind of noises influence the feature selection process and the above mentioned loss function fits with the Gaussian distribution noise density with maximal likelihood under the assumption for generation of feature subset. The loss function is re-casted to LP issues. After the elimination of noise, the feature subsets are generated from the provided data with reduced amount of redundant features. This can enhances the computational process more accurately and improves the accuracy.

Algorithm 1: L1 norm based SVR for DS

```
Input: available dataset = [X,Y]; X= input [array]; Y = c[array];
Output: prediction accuracy
Parameter initialization = variables ();
For learning available variables do
    Initialize the parameters
    For $i = 1$ do
        Decode the provided dataset
        Determine SVR parameter and feature subset
        Modify training set w.r.t. chosen feature subset to avoid redundant data
        Establish SVR classifier using parameter values to avoid Linear programming issues.
        Train SVR with training set
        Compute test set with trained SVR
        Determine overall fitness value
        Optimize SVR parameters and feature subset
        Establish SVR classifier with optimized parameter values
        Evaluate test set with trained SVR
    End for
End for
End
```

3. NUMERICAL RESULTS AND DISCUSSIONS

This section discusses the numerical results attained from after appropriate feature selection and applied using image processing toolbox and machine learning toolbox in MATAB. The selected features are tested for positive and negative DS from the facial images. The outcomes are reported based on sensitivity, specificity, accuracy, F1-score, precision and MCC. The mathematical expression of these aforementioned metrics is given as below:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$Specificity = \frac{TN}{TN + FP} \tag{11}$$

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{13}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (14)$$

Here, $TP, TN, FP,$ and FN are related to True Positive, True Negative, False Positive and False Negative respectively. The performance of L1-SVR is described in Table I and II. DS prediction rate is provided as 96.78% with 94.81% sensitivity and 98.50% specificity respectively.

Table I: Accuracy, sensitivity, specificity comparison

Approaches	Accuracy	Sensitivity	Specificity
L1-SVR	96.78%	94.81%	98.50%
Adam	95%	93%	92%
SGDM	91%	90%	92%
RMSPROP	95%	94%	94%
k-NN	71%	34%	92%
SVM	76%	59%	87%

The most appropriate features are chosen after performing L1-norm based SVR which assists in better understanding of DS identification. The multiple feature maps are given into the classifier for further classification process. The performance metrics of the anticipated L1-norm SVR is compared with existing approaches like Adam, SGDM, RMSPROPR, k-NN and SVM respectively. The accuracy, sensitivity, and specificity of L1-SVR o are 96.78%, 94.81%, and 98.50% respectively. The accuracy of L1-SVR is 1.78%, 5.78%, 1.78%, 26.78% and 20.78% higher than the prevailing approaches. The sensitivity of L1-SVR is 1.81%, 4.81%, 0.81%, 60.81% and 38.81% higher than the prevailing approaches. The specificity of L1-SVR is 8.5%, 8.5%, 4.5%, 8.5% and 11.5% higher than the prevailing approaches.

The precision of L1-SVR is 1.43%, 6.43%, 5.43%, 25.43% and 9.43% higher than the prevailing approaches. The F1-score of L1-SVR is 1.42%, 4.42%, 1.42%, 50.42% and 24.42% higher than the prevailing approaches. MCC of L1-SVR is 4.02%, 14.42%, 10.42%, 63.42% and 48.42% higher than the prevailing approaches. The features are chosen and the information are extracted from eyes, nose and mouth which provides effectual differences among DS patients from the normal patients. The similarities among the L1-SVR and other models demonstrate that the anticipated method successfully acquired the medical characteristics of DS patients. Fig 3 to Fig 9 shows the graphical representation of the metrics evaluation to project the performance of the anticipated L1-norm based SVR model. This model gives higher performance than the other approaches.

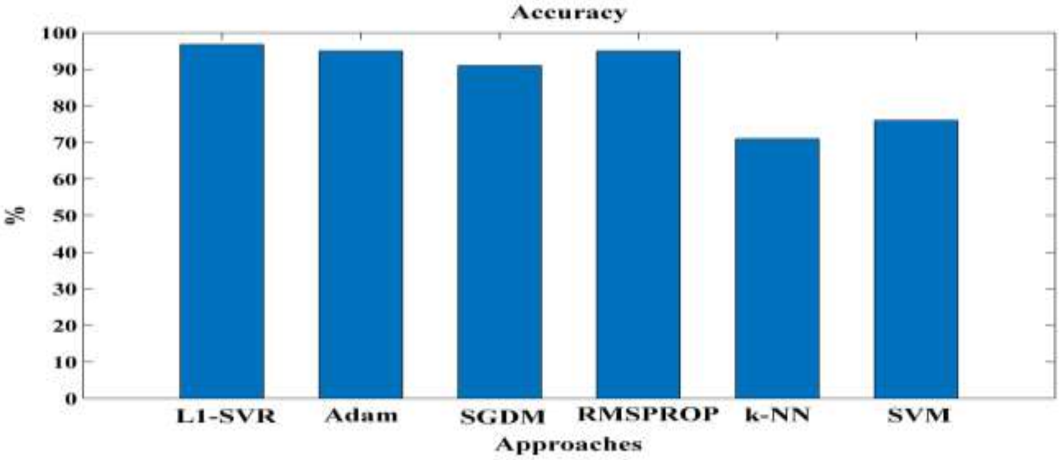

Fig 3: Accuracy computation

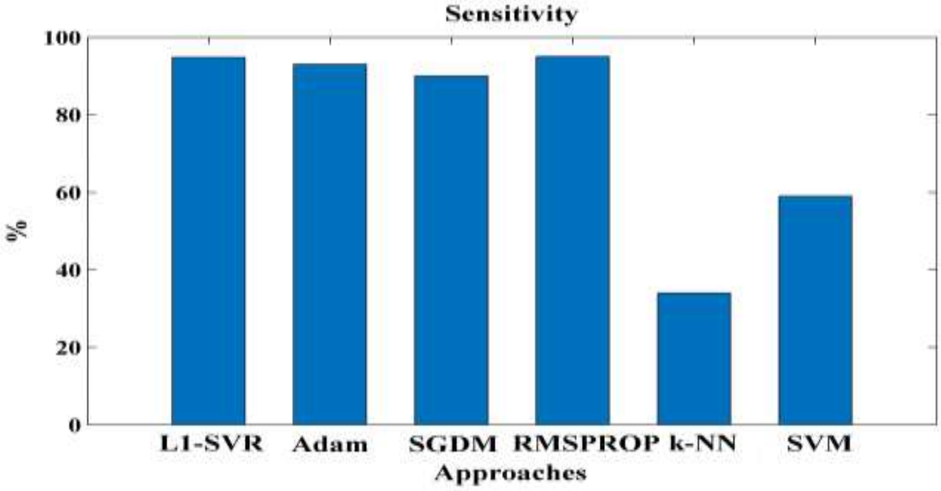

Fig 4: Sensitivity computation

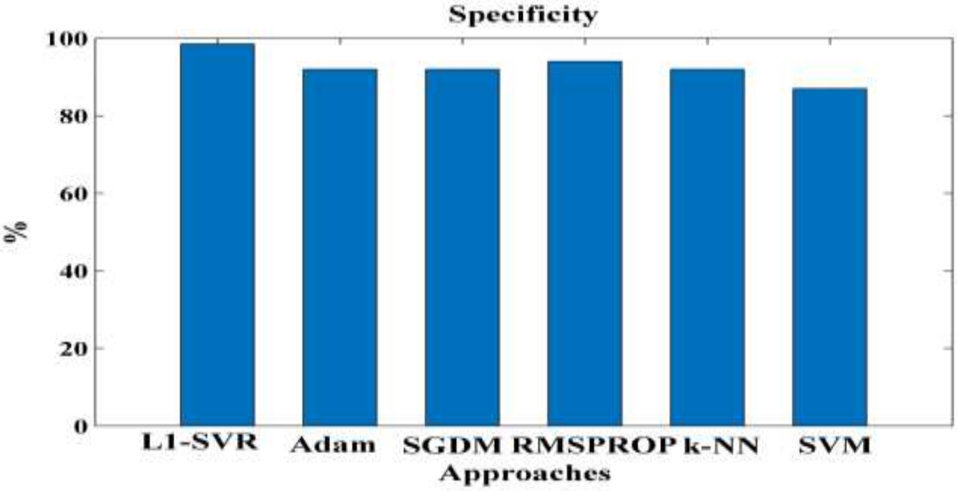

Fig 5: Specificity computation

Table II: Precision, F1-score and MCC comparison

Approaches	Precision	F1-score	MCC
L1-SVR	96.43%	96.42%	92.4%
Adam	95%	95%	82%
SGDM	90%	92%	86%
RMSPROP	91%	95%	89%
k-NN	71%	46%	33%
SVM	87%	72%	48%

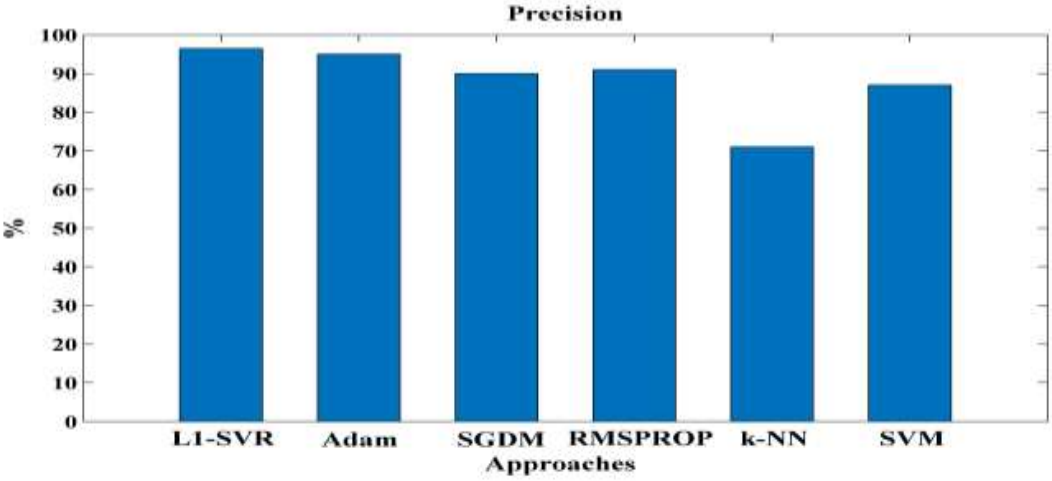

Fig 7: Precision computation

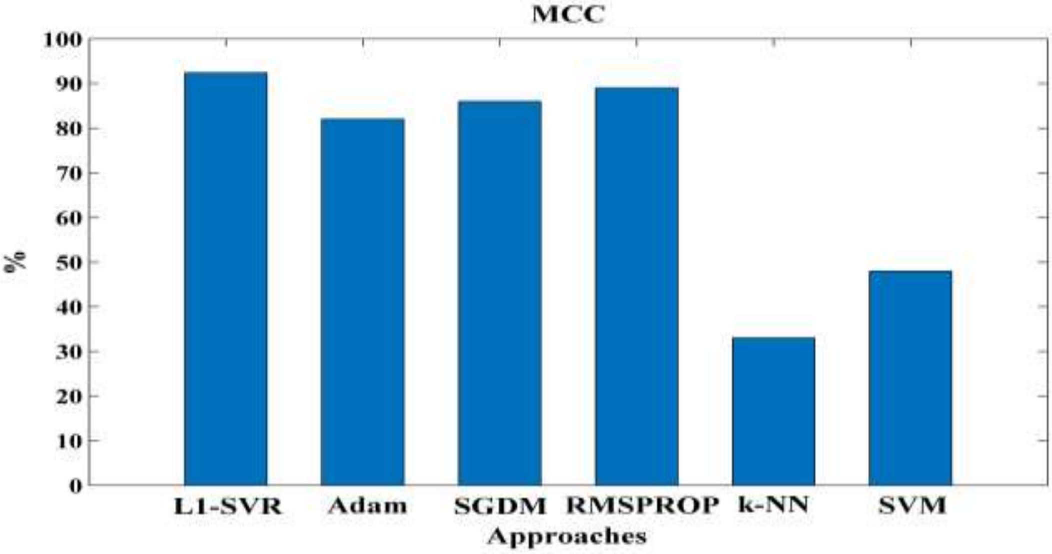

Fig 8: MCC computation

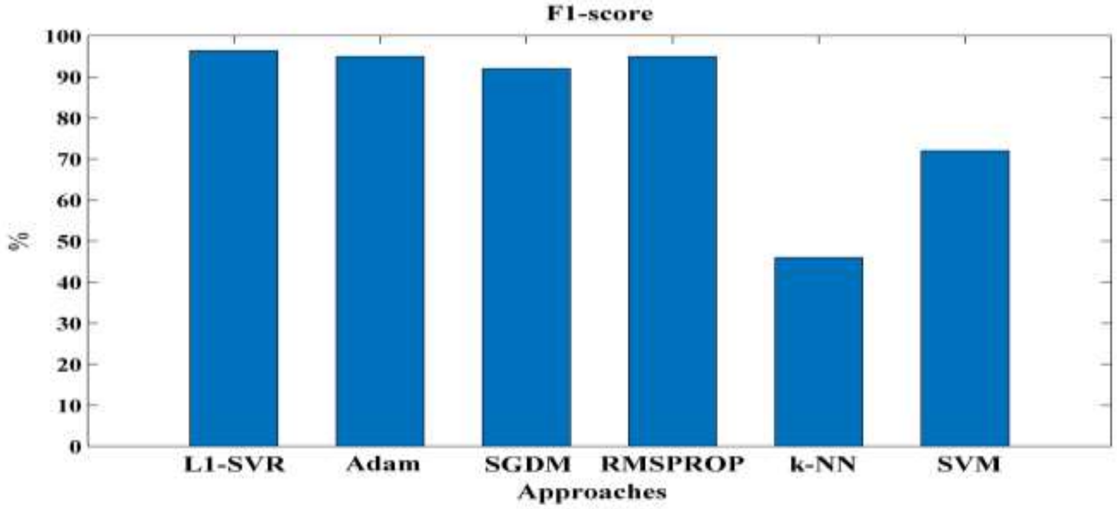

Fig 9: F1-score computation

The regions like eyes, mouth and nose during feature selection for DS identification has higher values than face recognition network. This specifies that the information attained from these regions is provided with higher consideration. These features were determined to be a most essential facial image characteristic of DS patients which is extracted from knowledge transfer learning. Without performing any pre-processing function, L1-SVR has to give more concentration towards facial information as the background occupies more space than faces themselves. However, the images have differences among the contrast and exposure which is given as the final performance outcomes.

4. CONCLUSION

DS is a common genetic disorder that influences human incidence with 1 of 700 from child birth. Due to its higher significance, prediction of DS is essential to avoid the complexities. Here, ML approaches are adopted to select the suitable features to enhance the performance of the system. This work employs L1-norm based SVR model to avoid noise at the dataset and to eliminate redundancy over the features. Thus, it forms a newer subset which is fed as an input to the classifier model. The proposed model work effectually than the existing approaches and gives better accuracy. The simulation is performed with MATLAB environment and the outcomes gives better trade-off among the others. The accuracy attained with this model is 96.78% which is higher than k-NN, SVM, Adam, SGDM, and RMSPROP respectively. In future, classification is done using ML approaches and an effectual benchmark dataset has to be constructed for enhancing the accuracy.

5. REFERENCES

[1] Vorravanpreecha, N.; Lertboonnum, T.; Rodjanadit, Studying Down syndrome recognition probabilities in Thai children with de-identified computer-aided facial analysis. Am. J. Med. Genet. A **2018**, 176, 1935–1940.

[2] Kruszka, P.; Porras, A.R.; Sobering, A.K.; Ikolo, F.A.; La Qua, S.; Shotelersuk, V.; Chung, B.H.; Mok, G.T.; Uwineza, A.; Mutesa, L.; et al. Down syndrome in diverse populations. Am. J. Med. Genet. A **2017**, 173, 42–53.

[3] Weijerman, M.E.; de Winter, J.P. Clinical practice. The care of children with Down syndrome. Eur. J. Pediatr. **2010**, 169, 1445–1452.

[4] Damasceno, L.N.; Basting, R.T. Facial analysis in Down's syndrome patients. RGO–Revista Gaúcha de Odontol. **2014**, 62, 7–12.

[5] Dimitriou, D.; Leonard, H.C.; Karmilo_-Smith, A.; Johnson, M.H.; Thomas, M.S. Atypical development of configural face recognition in children with autism, Down syndrome andWilliams syndrome. J. Intellect. Disabil. Res. **2015**, 59, 422–438.

[6] Chiu, R.W.; Akolekar, R.; Zheng, Y.W. Non-invasive prenatal assessment of trisomy 21 by multiplexed maternal plasma DNA sequencing: Large scale validity study. BMJ **2011**, 342, c7401.

[7] Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. Commun ACM **2017**, 60, 84–90.

[8] Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv **2014**, arXiv:1409.1556.

[9] Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **2017**, 39, 640–651.

[10] Yi, D.; Lei, Z.; Liao, S.; Li, S.Z. Learning Face Representation from Scratch. arXiv **2014**, arXiv:1411.7923.

[11] Hassner, T.; Harel, S.; Paz, E.; Enbar, R. E_ective Face Frontalization in Unconstrained Images. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4295–4304.

[12] Eidinger, E.; Enbar, R.; Hassner, T. Age and Gender Estimation of Unfiltered Faces. IEEE Trans. Inf. Forensics
Secur. **2014**, 9, 2170–2179.

[13] Davide Chicco. 2017. Ten quick tips for machine learning in computational biology. BioData Mining. 10, 35. doi: 10.1186/s13040-017-0155-3.

[14] Tara Eicher and Kaushik Sinha. 2017. A support vector machine approach to identfication of proteins relevant to learning in a mouse model of Down Syndrome. In Neural Networks (IJCNN), 2017 International Joint Conference on (May 2017), 3391-3398. IEEE.

[15] Dheeru Dua and Efi Karra Taniskidou. UCI Machine
Learning Repository. (2017). Retrieved Sep 20, 2017 from
https://archive.ics.uci.edu/ml/datasets/Mice+Protein + Expression

[16] Mao, J., Sun, Q., Wang, X., Muthu, B., & Sujatha Krishnamoorthy, S. (2020). The importance of public support in the implementation of green transportation in the smart cities. Journal of Computational Intelligence. Wiley publications .https://doi.org/10.1111/coin.12326. 26th April 2020

[17] Yasoda, K., Ponmagal, R.S., Bhuvaneshwari, K.S. K Venkatachalam, " Automatic detection and classification of EEG artifacts using fuzzy kernel SVM and wavelet ICA (WICA)" Soft Computing Journal (2020).