Minimax Theory and its Applications

Volume 11 (2026), No. 3s, 31 to 43

Hill Climbing Based Feature Selection and Self-Organizing Deep Auto-Encoder (Sodae) Classifier for Snps Dataset

N. S. Kalyan Chakravarthy[1], S. Jafar Ali Ibrahim[2*], Madhala Srinivasulu[3], Gopinath Veerisetty[4], Raenu Kolandaisamy[5], Kousik Nalliyanna Goundar Veerappan[6]

[1,2]Institute of Computer Science and Digital Innovation, UCSI University, Jalan UCSI, UCSI Heights (Taman Connaught), Cheras, Kuala Lumpur, Malaysia, Email: SuryaKalyan@ucsiuniversity.edu.my[1], JafarAli@ucsiuniversity.edu.my[2]

[3,4] Center for AI and ML, QIS College of Engineering and Technology, Ongole – 523272, AP, India, Email: srinivasulu.m@qiscet.edu.in[3], veerisetty.gopinath@qiscet.edu.in[4]

[5]School of Technology, Asia Pacific University of Technology and Innovation(APU), Jalan Teknologi 5, Taman Teknologi, Kuala Lumpur, 57000, Malaysia, Email: Raenu.k@apu.edu.my

[6]Department of Computing and Technology, Arden University, London, UK, Email: kngveerappan@arden.ac.uk

*Correspondence to: S. Jafar Ali Ibrahim (JafarAli@ucsiuniversity.edu.my)

Abstract

Single Nucleotide Polymorphisms (SNPs), are known as "snips" in the genetics community, are the most common kind of genetic variation in humans.A single nucleotide might vary in each SNP, which is a fundamental component of DNA. Significant SNPs in many human illnesses have been found using a variety of Machine Learning methods.Issues in the area of SNP identification include the dimensionality curse and the possibility of uneven numbers of healthy and patient samples. Thus the feature selection and classification has been efficient solution to solve these challenges in the SNP dataset. A novel method is presented in this research to feature selection, and classification of two different SNP datasets. The nominal SNP data are first transformed the mean encoding to numerical data.For feature selection, the Hill Climbing (HC) algorithm was then devised, which eliminates characteristics that are unnecessary and redundant.In order for the classification to create its structure based on input data, a Self-Organizing Deep Auto-Encoder (SODAE) is used. To assess a classifier's performance, Breast cancer and thyroid cancer SNP datasets have been gathered from the Gene Expression Omnibus (GEO). Precision, recall, F-measure, and accuracy for classification techniques are used to evaluate the effectiveness of the findings. When compared to techniques like k Nearest Neighbor (kNN), support vector machines (SVM), and support vector machines recursive feature elimination (SVM-RFE), the findings show that it has been highly successful.

Keywords: Deoxyribonucleic acid (DNA), Singular Nucleotide Polymorphism (SNP), Genomic Analysis, Gene Expression Omnibus (GEO), Feature selection, classification, Hill climbing, and Self-Organizing Deep Auto-Encoder (SODAE).

1. Introduction

Deoxyribonucleic acid (DNA) is a substance that contains the genetic information for an organism. Everything needed to create and sustain a creature is encoded in DNA.During reproduction, genes—segments of DNA that contain genetic information—are passed from parents to children [1, 2]. The most prevalent kind of DNA mutation is known as a single nucleotide polymorphism (SNP). SNPs may have a substantial role in the relationship between complicated disorders, according to studies [3, 4]. To study complicated diseases, a number of machine learning techniques have been developed.Naive Bayes (NB), kNN, Random Forest (RF), Dimensionality Reduction (DR), Gradient Boosting (GB), AdaBoosting (AB), Linear regression (LR), Logistic Regression (LR), Decision Tree (DT), and SVM are a few of these[5,6].

Feature Selection (FS) is a technique that uses only relevant data to reduce the input variable to model.Choosing suitable characteristics for machine learning automatically depending on the kind of issue is what this approach entails [7]. There is a wide variety of FS algorithms, including filter, wrapper, and embedding, and each

approach has the potential to be appropriate for a certain kind of data[8].Without regard to the sort of classifier that is being used, the filter technique chooses the feature.This technique has the benefit of being easy and irrespective of classifier type.As a result, only one round of feature selection is necessary, and it disregards feature dependencies, interactions with the classifier, and finally, individual feature consideration [9]. In wrapper, selection of features is done by considering it as a search problem, and different combinations are made. It trains the algorithm by using the subset of features iteratively. In embedded algorithms, both the filter and wrapper approaches' benefits are combined by considering the interaction of features with lower computational cost.

Among these methods, filter and wrapped approach techniques are combined together for FS. The suggested work uses a Hill climbing (HC) method to lower the complexity of the classification job, while simultaneously accounting for the original relevance of each attribute in the database, a random mutation technique.The search is conducted using the approach of random mutation and the original relevance of each attribute in the classification process being taken into consideration at all times[10]. The results of the algorithms give best specific SNP data and give with great accuracy, compare the patient and healthy samples. The classification job is used to identify significant SNPs, and the FS method is used to determine which SNP properties are crucial. As a result, the FS algorithm improves the classifier's performance. After the FS, then classification takes major important role for improving the disease diagnosis results.

Data science, which also covers statistics and predictive modeling, is fundamentally dependent on deep learning. Deep learning has the potential to significantly speed up and simplify the work of data analysis for data scientists who are entrusted with gathering, analyzing, and interpreting large volumes of data. In this content, auto-encoder based deep learning algorithm has been focused for classification [11]. They are trained by using the use of massive volumes of data that have been tagged as well as neural network designs that learn characteristics directly from the data. In order to generate a data representation that is more abstract, these approaches are built by combining a number of different nonlinear mappings [12]. Unsupervised techniques are used to make a deep representation by deep model. Many different applications have been successful using Deep Learning approaches.

In this paper, HC algorithm is introduced for feature selection algorithm. The HC technique chooses a minimal feature number of nodes and layers and a feature subset with excellent classification performance.The classification process uses the Self-organizing Deep Auto-Encoder (SODAE) method, which autonomously builds its structure from incoming data.To evaluate the proposed technique, the SNP database obtained from the Gene Expression Omnibus (GEO) database was employed.Comparing the outcomes to previous efforts, the findings show that it has been very effective.

2. Literature Review

Nomin Batnyam et al [13] proposed to the data structure and the huge feature space, a Feature Selection Distance Discriminant (FSDD), as well as a Relief and R-value based Feature Selection (RFS), and a Clearness Based feature selection (CBFS), may be quickly accomplished and connected with complicated illnesses. SNP analysis is used in order to ease the categorization of SNP data. Raid Alzubi et al [14] based on the combination of a filter- and wrapper-based methodology, a hybrid feature selection strategy for complicated illnesses has been developed.SNPs are evaluated using a variety of feature selection strategies, including Minimum Redundancy Maximum Relevancy (mRMR), Conditional Mutual Information Maximization (CMIM), and Relief. Four classifiers—SVM, NB, Linear Discriminant Analysis (LDA), and kNN—have been evaluated on five distinct SNP data sets. When compared to alternative feature selection approaches, the findings demonstrate that the feature selection methodology is effective.

Shah and kusiak[15] suggested a novel method for forecasting pharmacological efficacy based on data mining (DM) and genetic algorithm (GA).A Global Search Mechanism, a Weighted Decision Tree (WDT), a Decision-Tree-Based Wrapper (DTBW), a Correlation-Based Heuristic (CBH), and the technique of finding overlapping feature sets to choose relevant genes.Their objective was to identify the gene/SNP patterns that affect medical development. The number of features has been reduced by 85% as a direct consequence of the feature selection technique. The relevant gene/SNP set's cross-validation specificity and accuracy both increased by 3.2% and 10%, respectively. Even though there are less features, the expertise level has improved.

Podder et al [16] The currently available genotype calling method by manually looking at the Array Primer Extension (APEX) data, is capable of calling single SNP genotypes.A highly parallel microarray platform and the recommended Linear Discriminant Analysis (LDA) approach make a potent combo. The proposed method is fully-automated genotyping, based upon using dynamic variable selection.

Silva et al [17] designed a new algorithm RF and Cluster Analysis (CA) integration for the database used to suspect illness. Multiple cluster analyses using the integrated framework were performed on (1) SNPs and (2) SNPs with the highest feature relevance scores obtained using random forest. Trait-association tests were run on the SNP-sets that appeared after cluster analysis. The final findings improve our knowledge of the causes of illness and offer us information to help patients get a more accurate evaluation of their disease risk.

Jiao et al [18] Autism Spectrum Disorder (ASD) may be predicted using a model created using SNPs.Children with ASD are categorised according to the Childhood Autism Rating Scale (CARS) as mild, moderate, or severe. In order to develop prediction models, we used the following three machine learning strategies for each kid that was segmented according to SNPs and genes associated with ASD: (1) Impasses in Decision Stumps (DSs) (2) Alternating decision trees (AD Trees) to choose from (3) Stretching Trees. With accuracy of 67%, sensitivity = 0.88, and specificity = 0.42, Decision Stumps and Flex Tree provide superior classifiers. The final findings accurately classify the severity of ASD symptoms. Begum et al [19] Decision-making, illness detection, and improved patient care at lower costs all depend heavily on the suggested data mining methods of NB, DT, Multilayer Perceptron (MLP), and Radial Basis Function Network (RFN).The Thyroid Stimulating Hormone (TSH), T3, and T4 correlate with hyperthyroidism and hypothyroidism, and these classification systems may also identify the association between TSH, T3, and T4 and gender in relation to these conditions. The results show that all categorization models have a substantial level of accuracy.

Nunes et al [20] to choose a subset of characteristics for character recognition, an Optimized Hill-Climbing algorithm (OHC) has been presented.Running the MLP classifier results in a decrease in the number of initial characteristics without a significant impact on recognition rates.The experiment demonstrated that some loss in recognition rate should be taken into account when choosing a feature subset. To simplify the classification work on digital recognition, complexity is reduced by 86.54% using a cascade classifier with different-sized feature subsets.Without the need of complicated and costly hardware designs, the suggested OHC technique has shown the implementation of the wrapper approach. Li [21] Exhaustive Searching based on Multi-Objective Optimization (ESMO), a new rapid scheme approach known was presented.When ESMO evaluates epistatic SNP combinations to see if they are consistent with heterogeneity complicated illnesses, Mutual Entropy (ME) and Bayesian Network (BN) techniques are combined.Time complexity is increased by designing non-dominant sort and the top k selection procedure are the foundation of an adaptive framework.The result demonstrates the ability to efficiently identify heterogeneity of complicated population structures as well as promptly address epistasis.

Jing and Shen[22]proposed Multi-Objective Ant Colony Optimization Epistasis Detection (MACOED), a Multi-Objective Heuristic Optimization (MHO), was compared to both LR and BN approaches for identifying genetic interaction.MACOED algorithms are able to tackle the space and temporal complexity of high-dimensional issues. It has been shown that it is complimentary, which produces a greater power with a reduced false-positive rate, and it is computationally feasible for use with huge databases.Li et al [23] suggested a three-stage deep learning architecture with a merging model that includes: (1) the detection of epistasis; (2) the formation of clusters; and (3) the prediction, using a technique based on deep learning, of epistasis and heterogeneity in complicated disorders. In order to identify many possible diverse kinds of complex illnesses are influenced by groups of epistatic SNPs, the epistasis identification step uses a multi-objective optimization strategy.The k-mean to categories the case group's subtypes, a clustering technique is utilized. Finally, a disease prediction deep learning model based on graphics processing unit has been developed (GPU). The final findings on pure and heterogeneous databases demonstrate the potential usefulness of the technique, and it has the ability to serve as an alternative to previous approaches.

Hess et al [24] suggested Using a Sparse Regression (SR) approach, the joint distribution of SNPs was coarsely screened, and the SNP partitions that were discovered during the screening were used to train a large number of Deep Boltzmann Machines (DBMs). This was followed by analyzing the results of the screening.Combining statistical analysis and sparse regression, The DBMs are used to retrieve aggregate characteristics that represent SNP patterns and the relevant SNPs.It demonstrates how to find complicated SNP patterns in simulated case-control data and improve outcomes from univariate methods.Applications assess time-to-event endpoints; the suggested method found three SNPs in a validation sample that seem to collectively impact survival.As a result, partitioned learning of DBMs becomes a fascinating supplementary method for evaluating SNP data, demonstrating the advantages of simultaneously exploring SNPs over traditional univariate studies.

Boutorh and Guessoum [25] a novel combining grammatical evolution (GE) and evolutionary algorithms with association rule mining (ARM) and neural networks (NN). In rule extraction and GA optimization, it is also known as Genetic Algorithms- Neural Networks- Grammatical Evolution- Association Rule Mining (GA-NN-GE-ARM).GE has developed ARM to reduce dimensionality by extracting correlations between SNPs

concurrently from two independent datasets of case and control samples.A NN is used for effective categorization, on the other hand. GA sets the settings for the two combined procedures.Four separate SNP datasets from the National NCBI GEO website have been used to test the suggested technique.When integrated with several classifiers, the final results achieve higher than average categorization accuracy a number of feature selection strategies.

3. Proposed Methodology

There are three steps to the proposed process: (A) A pre-processing stage includes nominal SNP data encoding and the removal or replacement of missing values from SNP data. (B) A step known as feature selection, when a good FS algorithm selects relevant SNPs. (C) At this level, which is a classification stage, SNP data is categorized using self-organizing auto-encoders.For assessing the effectiveness of the classifiers, classification measures including Precision, Recall, F-measure, and accuracy have been established. Figure 1 displays the whole procedure.



Figure 1. General Process Of The Proposed System

3.1 SNP dataset

The NCBI GEO website served as the source for all of the SNP data analyzed in this investigation. Other genetic data with Next Generation Sequencing (NGS) are archived and publicly distributed via the GEO public

repository.The suggested technique is used to study complicated illness using the GSE67047 [26] and GSE16619 [27] samples, each of which has 500,000 SNP, with 111 samples altogether, 69 from cases and 42 from controls.With different numbers of samples (case and control), characteristics (SNPs), and genetic patterns, the SNP databases for thyroid cancer and breast cancer include these two diseases.Current SNP research based on artificial intelligence (AI) typically uses the five SNP datasets. Additionally, one of the most dependable databases for biology in the world, GEO of the National Center for Biotechnology Information, hosts these datasets.

3.2 Pre-processing

Three phases make up the pre-processing phase. Prior to replacing No Call or missing data with appropriate quantities in the second stage, duplicate SNPs are first identified and eliminated.

Redundant features: Since these characteristics or SNPs are unable to distinguish between the two groups, they are referred to as redundant features. Case and control samples with identical SNP values are also referred to as redundant samples. It will not be useful for feature selection and classification, if BB value is recorded for a particular SNP in all case and control scenarios.

Missing values: In this research, in certain SNP datasets, the mode of feature (also known as the feature type), it is used to replace the No Call value with an approximated value and is regarded as a new feature type in certain instances. Each SNP that contains more than 10% of No Call values is removed from the analysis.

Encoding method: Although there are no nominal SNP data in this research, encoding the SNP data into a numerical quantity requires an appropriate encoding technique[28]. The target label may be taken into account throughout the mean encoding technique's encoding process, in similar to other encoding techniques, relationship to the goal. This makes the mean encoding method intelligent. Mean encoding applies equation to get the numerical value for the special nominal characteristic (1).

$$Mean\ encoding\ of feature\ i = \frac{Number\ of\ true\ targets\ under\ the\ feature\ i}{Total\ number\ of\ targets\ under\ the\ feature\ i} \tag{1}$$

3.3 Feature Selection by Hill Climbing

Feature Selection (FS) attempts a selection of attributes from the incoming data to be selected so that it may deliver accurate results for classification and prediction. A measure of similarity is used in the computation of the redundancy.Numerous metrics of similarity have been devised, with the Correlation Coefficient (CC), as indicated in equation, being one of the most well-known measurements (2).

$$\rho(x_i, x_j) = \frac{\sum_{k=1}^{N}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{Var(x_i)Var(x_j)}} \tag{2}$$

Other similarity metrics exist as well, such as the Symmetrical Uncertainty (SU) and Maximal Information Compression Index (MICI), as illustrated in equations (3) and (2) (4),

$$2\lambda(x_i, x_j) = Var(x_i) + Var(x_j) - \sqrt{\left(Var(x_i) + Var(x_j)\right)^2 - 4Var(x_i)Var(x_j)\left(1 - \rho(x_i, x_j)^2\right)} \tag{3}$$

$$SU(x_i, x_j) = \frac{H(x_i) - H(x_i|x_j)}{H(x_i) + H(x_j)} \tag{4}$$

Given that this metric has also been successful with sparse data [30], The redundancy between two features equations (5) is calculated using absolute cosine similarity, as reported in this article.

$$\left|Cos(\theta_{x_i,x_j})\right| = \left|\frac{\langle x_i, x_j\rangle}{\|x_i\|.\|x_j\|}\right| \tag{5}$$

$\langle .,. \rangle$is referred to as the internal product $\|.\|$ is the Euclidean norm. It will be as shown in equation (5)$0 < |Cos(\theta_{x_i,x_j})| < 1$ orthogonal and collinear, respectively, describe the two properties, with 0 and 1 as their corresponding values. Figure 2 depicts the introduction of the five-stage Hill Climbing (HC) feature selection method.Feature vectors from a database with two SNPs are included in the database collection that makes up the

first step. Training and testing are assessed in the second step while taking the initial input SNPs data into consideration.The HC method is suggested to verify data in the third stage in order to identify several candidates for feature subsets. Depending on how important it is for classification accuracy, a randomly chosen characteristic is eliminated or left in.In the fourth step, the chosen subset database is ultimately assessed using useful database. Finally, the datasets that have been modified to match the feature subset are retrained using the subset of features that have been chosen. The structure's flow is seen in Figure 2.

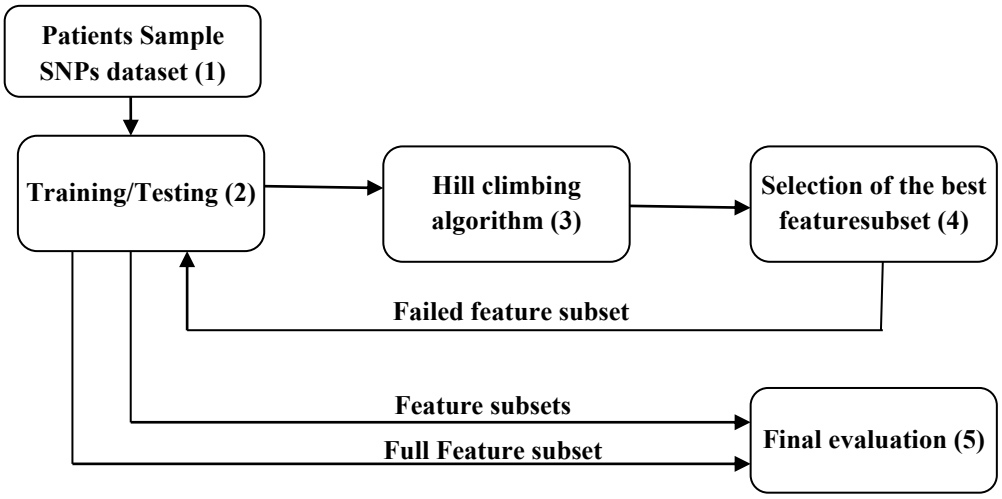


Figure 2. Overview Of The Proposed Method

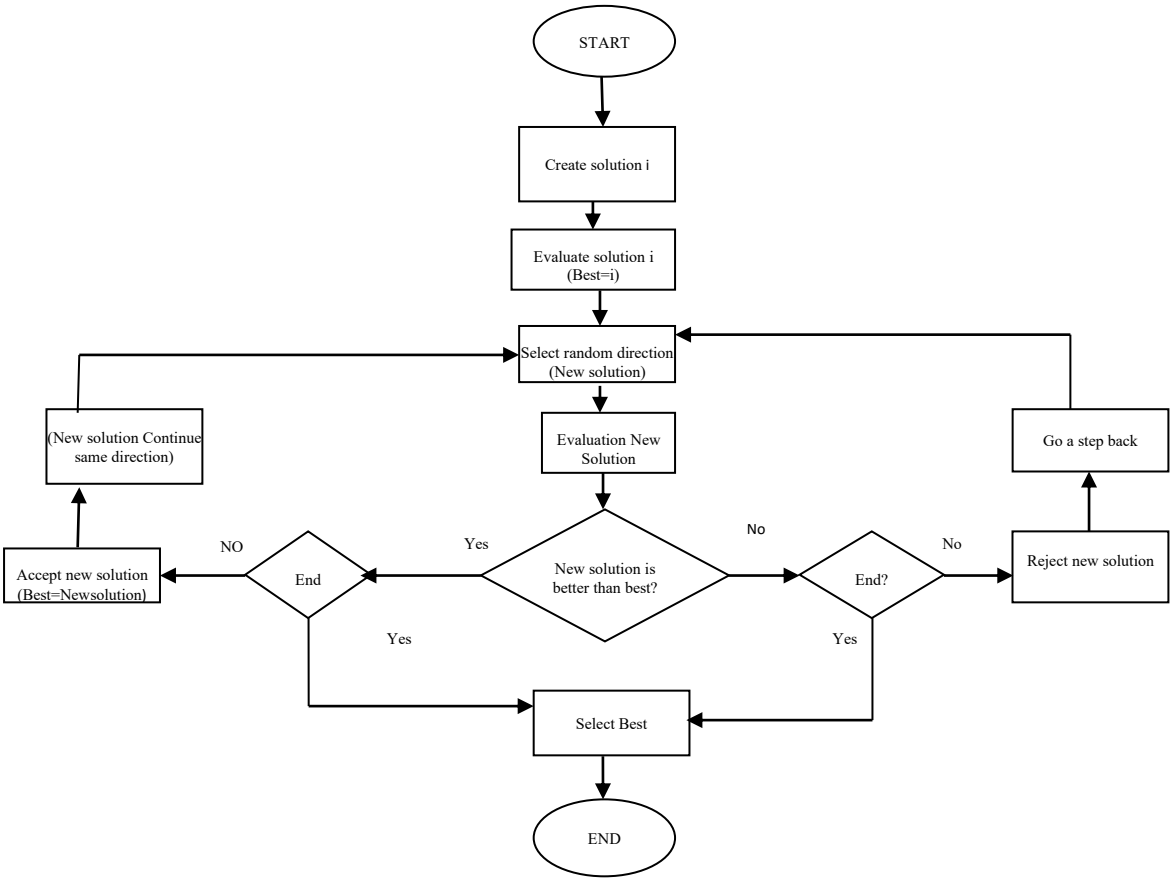


Figure 3. Flowchart of the Hill Climbing

Figure 2 provides a basic flowchart of the proposed system. Initially it starts with normalized SNPs dataset. Initializes and analyses the dataset's total number of features. After that randomly selects new features and evaluates them. If the new features attained higher results kept it is best features else go to new evaluation.

Complete all the features until it reaches maximum iteration. Including the whole feature set. The computational complexity of the classification task, however, was shown to have significantly decreased.Complexity of the original classifier is discussed in equation (6),

$$compl_{oc} = ns \times nc \tag{6}$$

$$compl_{cc} = \sum_{i=1}^{NC} (ns_i \times nc_i) \tag{7}$$

ALGORITHM 1: Hill Climbing algorithm
Step 1: If the starting state is the desired state, determine if it was successful, and if so, quit.
Step 2: Loop until a fix is identified or no more operators are available to use.
Step 3: To the present state, choose and use an operator.
Step 4: Check new state:
1. Return success and stop if it's a target state.
2. A new state should then be designated as the current state if it is superior than the alternative.
3. Otherwise, if the situation is not improved, go back to step 2.
Step 5: Exit.

In algorithm 1, the starting state is initialized in the first phase, followed by a neighbor state evaluation.The algorithm pauses and returns success if it has a large cost, in which case it moves on to the adjacent state. The current state is taken to be the beginning state if it is not otherwise.Once the solution state is reached, go to step two and repeat the same process. The third step involves assuming that the original state is the current state and once again comparing it to the surrounding state.The algorithm finishes after step four and provides success if the new current state matches the intended or state.If else is not the case, it selects a new neighbor as its current state, compares it once again with neighboring states, and then continues.Once the desired state or solution has been reached, the iterative process is repeated. Finally, success and leave from the return.

3.4 Sodae

This work proposes a unique kind of AE called the SODAE, which may choose its structure depending on input data.The SODAE builds its structure in two stages by first determining the amount of nodes and layers it will need.The quantity of inputs defines the number of nodes in the layer after AE converts input features into a new feature space, equation states that the first step occurs when the size of the encoding space is larger than the input space (8). $n_{initial}$, d_{input}andα, the input dimension, a user-defined coefficient, and the initial number of nodes in the encoding layer are listed in that sequence, and nodes. Also, $\alpha$is regarded as having several ($1 < \alpha$) thus $n_{initial} > d_{input}$.

$$n_{initial} = \alpha \times d_{input} \tag{8}$$

Train AE with $n_{initial}$with training data, the encoding layer's nodes. Then, at the encoding layer, the Relief-F method is used to assess the produced features and determine the weight for each feature. Figure 4 illustrates the characteristics that are maintained and those that are eliminated based on whether their weights above the threshold (user-defined parameter).

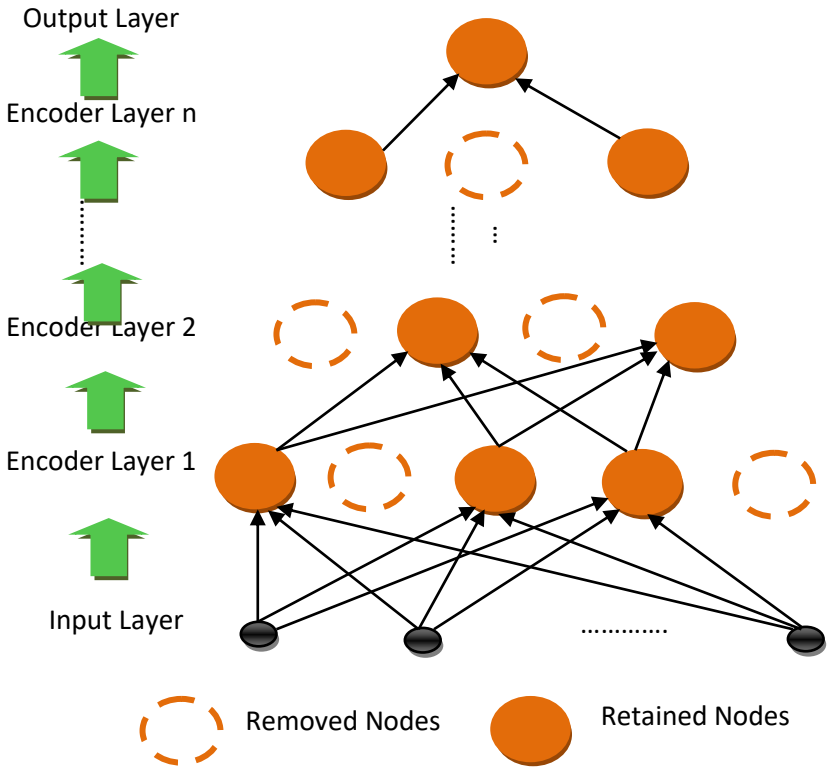


Figure 4. Sodae Classifier

According to classification, the encoding layer used the sigmoid activation function, and the fine-tuning layer makes use of Linear with Soft-max.A variable learning rate was determined based on the theory of cyclical learning rates [28]. The user may choose the width and depth of the self-organizing deep auto-depth encoder in the suggested approach by using the appropriate settings for α=1.1, maximum number of AEs, number of neighbors = 10, and threshold = 3.

4. Results And Discussion

In a brief, the following procedures were carried out for each and every set of SNP data, and the results of the simulations are presented in this part of the article. The existing classifiers and proposed system is implemented via the MATLAB simulation.

4.1 dataset

GSE67047- Thyroid cancer,GSE16619- Using the Affymetrix 250K Nsp and 250K Sty microarrays or the Affymetrix SNP5.0 microarray, we assessed copy number variation in 161 original samples of breast cancer to find new gene amplification events that may contribute to the development of data sets.DNA taken from primary samples was used in Affymetrix SNP arrays in accordance with the manufacturer's instructions.From the signal strengths of the probe sets, copy number variations were calculated. Utilizing precision measurements, the classifiers' performance has been evaluated, recall, F-measure, and accuracy.

4.2 Evaluation Measures

The categorization using the SNP data produced a result that represents the degree to which the suggested approach is able to differentiate between healthy and sick groups on the basis of the chosen SNPs.The results of proposed classifier and existing classifiers was evaluated via four measures like precision, recall, F-measure, and accuracy It, as demonstrated in equations(9-12), shows how the suggested strategy might be used to forecast instances.

$$\text{Precision} = TP/\ TP + FP \tag{9}$$

$$\text{Recall} = TP/\ TP + FN \tag{10}$$

$$F - measure = 2(Precision \times Recall / Precision + Recall) \quad (11)$$

$$Accuracy = (TP + TN / TP + FN + FP + TN) * 100 \quad (12)$$

The numbers for each kind of outcome are denoted by the words True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN), respectively.The comparison of the proposed technique with other methods shows that the proposed FS method and suggested SODAE classifier were the most accurate in selecting and classifying SNPs(See Table 1).

Table 1. Performance Of Classifiers With Fs Methods On Snps Dataset

Metrics	Feature selection/ Classifiers	kNN	SVM	SVM-RFE	SODAE
Precision (%)	Relief-F	69.27	82.13	84.13	87.15
	FCBF	75.42	88.15	90.15	91.67
	CMIM	84.79	90.48	91.37	92.79
	Hill climbing	90.19	91.78	92.74	95.52
Recall (%)	Relief-F	68.47	86.54	88.09	91.57
	FCBF	74.52	91.15	92.19	93.79
	CMIM	87.49	93.38	94.57	94.77
	Hill climbing	91.95	94.86	95.72	96.12
F-measure (%)	Relief-F	66.42	85.34	86.14	88.36
	FCBF	72.47	90.15	91.17	92.73
	CMIM	86.14	91.43	92.47	93.78
	Hill climbing	90.57	92.32	93.73	94.82
Accuracy (%)	Relief-F	67.15	85.17	87.15	88.41
	FCBF	74.41	86.18	89.63	90.15
	CMIM	84.13	87.71	90.81	91.43
	Hill climbing	88.14	91.81	93.39	95.18

Comparisons were made between the outcomes of the suggested approach (Hill Climbing + SODAE) and those of other extensively used FS methods in this area.The Relief-F [29], CMIM and Fast Correlation-Based Feature Selection (FCBF) [30] are the recognized FS approaches, and the top 100 SNPs were chosen based on the results of applying these methods.The chosen SNPs were also evaluated using well-known classifiers like kNN, SVM, Support Vector Machine-Recursive Feature Elimination (SVM-RFE), and SODAE. The efficiency and F-measure of the aforementioned combination approaches for SNP gene database were then calculated.

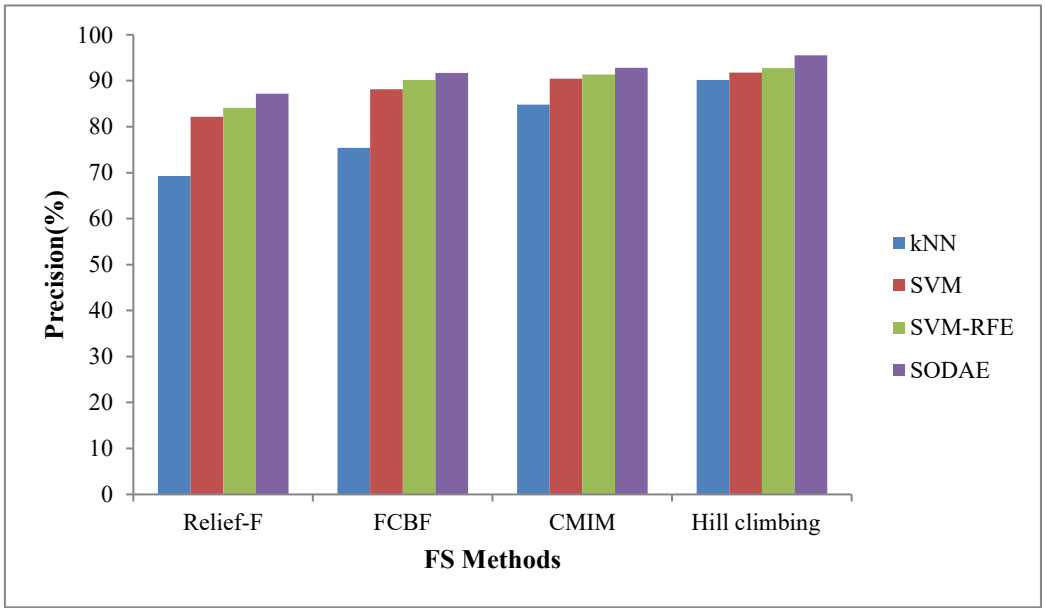


Figure 5. Precision Comparison Of Fs Methods Vs. Classifiers

Figure 5 compares the precision outcomes of the proposed technique (HC+SODAE) with those obtained using other widely used Feature Selection (FS) methods.Higher precision findings of 95.52% are provided by the suggested HC+SODAE classifier, whereas other methods such as kNN,SVM, SVM-RFE and SODAE also gives higher precision for proposed HC based FS algorithm such as 71.18%,90.19%, 91.78%, and respectively for gene dataset (Refer Table 1). The proposed FS algorithm gives highest results for all classifiers than the existing FS methods. Since the proposed work, filter and wrapper based FS algorithm is performed via the intelligent method.

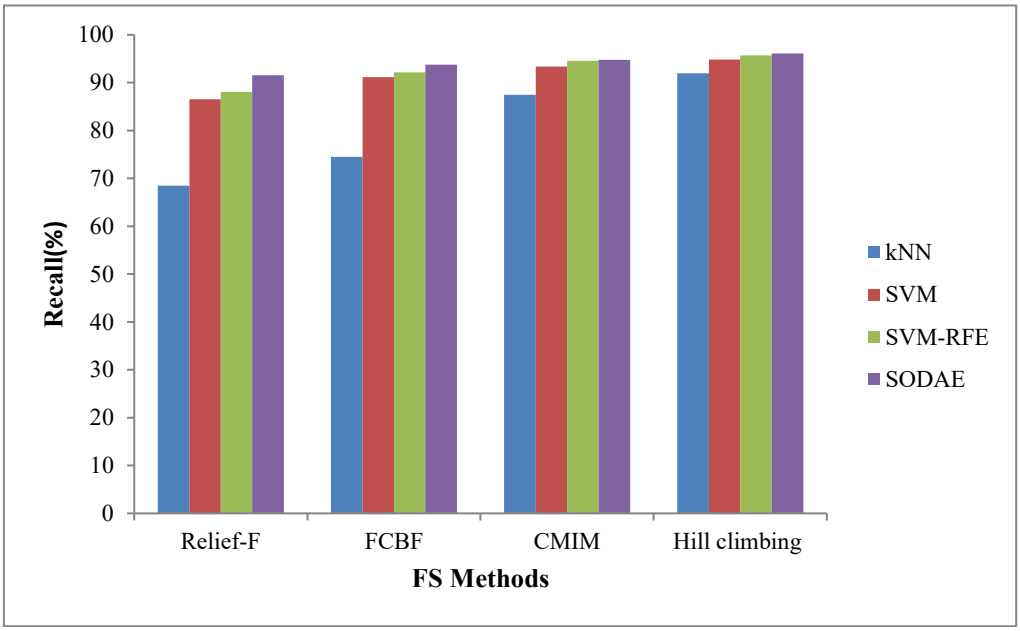


Figure 6. Recall Comparison Of Fs Methods Vs. Classifiers

Figure 6 shows the recall results after comparing the suggested approach (HC+SODAE) with other widely used FS methods. Recall results give 96.12% where compared with HC+SODAE, other methods such as kNN, SVM, SVM-RFE and SODAE. The proposed HC based FS algorithm such as 74.46%, 91.95%, and 94.86%, respectively for SNP dataset (Refer Table 1). The proposed FS algorithm gives highest results for all classifiers than the existing FS methods.

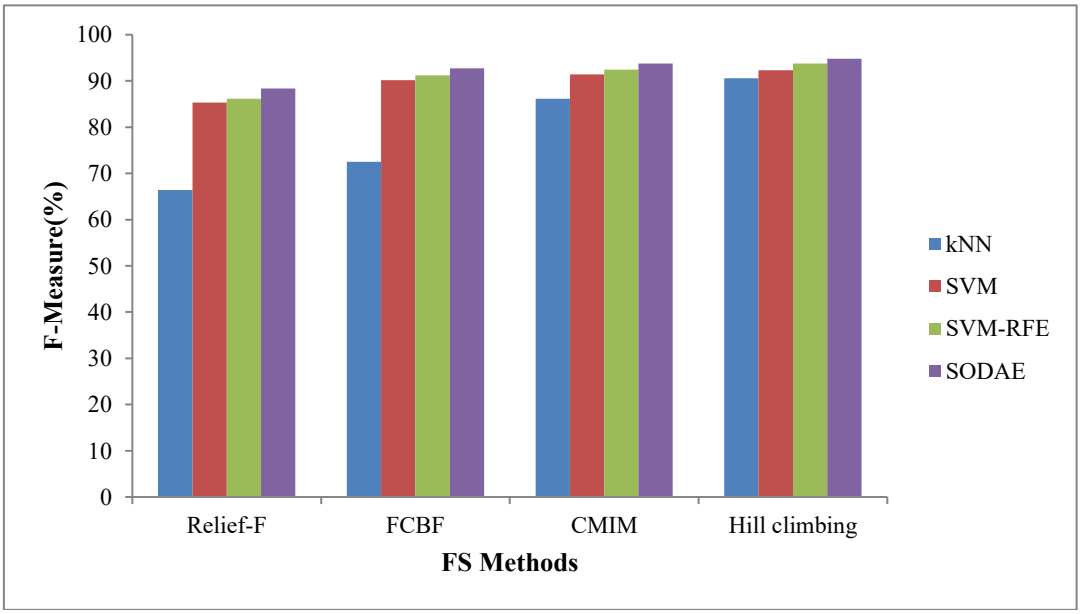


Figure 7. F-Measure Comparison Of Fs Methods Vs. Classifiers

Proposed method (HC+SODAE) was compared with other FS methods via Figure 7 discusses classifiers using F-measure. Comparing with other methods such as kNN, SVM, SVM-RFE andSODAE also gives improved F-measure value of 94.82% HC+with the SODAE. The proposed HC based FS algorithm such as 72.82%, 90.57%, and 92.32% respectively for SNP dataset (Refer Table 1).

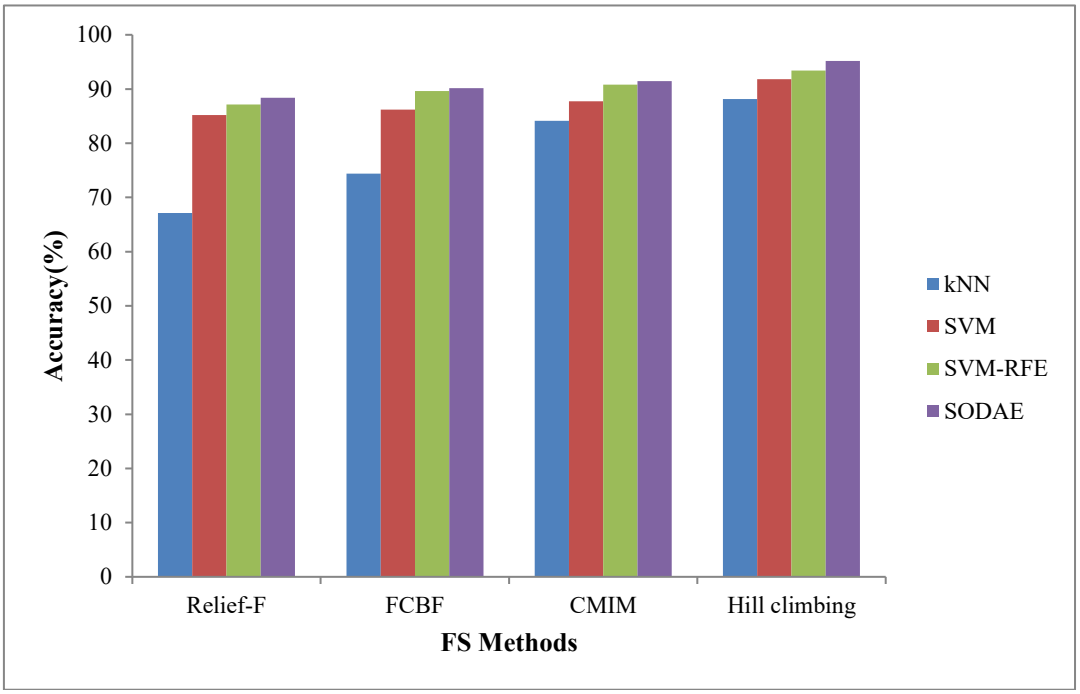


Figure 8. Accuracy Comparison Of Fs Methods Vs. Classifiers

The accuracy results in relation to FS approaches using classifiers are shown in Figure 8 in a performance comparison. The highest accuracy results of 95.18% combined with HC+SODAE, whereas other methods such askNN, SVM, SVM-RFE, and SODAE. The improved accuracy value for proposed HC based FS algorithm such as 90.15%, 88.14%, and 91.81% and respectively for genedataset (Refer Table 1).

5. Conclusion And Future Work

An innovative strategy for classifying important SNPs in complex disease has been developed in this respect.Data from SNPs were converted from nominal to numeric using the mean encoding approach.The Hill Climbing (HC) approach, which incorporates the Correlation Coefficient (CC), was used to identify important SNPs as follows, and cosine similarity as fitness function. HChas shown to be an interesting strategy by both filter and wrapper-based method with lower computational time.In order to identify a relevant feature's structure, SODAE is used. This eliminates the need to test different structures with varying numbers of nodes and layers throughout the learning process.In each and every one of the trials, SODAE was constructed automatically in accordance with the SNP data, and HC has achieved the best level of accuracy in terms of minimizing the amount of time and computing cost associated with high-dimensional SNP data.Statistical metrics were used to assess the potential of the classifiers in the proposed technique when it was applied to the Thyroid cancer and Breast cancer SNP datasets.In future work, avoid the duplication of the SNP's detection and healthy patients.

References

1. Divya, S., Bhavani, Y. and Thota, M.K., 2022. A Survey on Genomic Dataset for Predicting the DNA Abnormalities Using Ml. International Journal of Advanced Computer Science and Applications, vol. 13, no. 5, pp.311-319.

2. Yousefi, S., Abbassi-Daloii, T., Kraaijenbrink, T., Vermaat, M., Mei, H., van't Hof, P., van Iterson, M., Zhernakova, D.V., Claringbould, A., Franke, L. and t Hart, L.M., 2018. A SNP Panel for Identification of DNA and RNA Specimens. BMC genomics, vol.19, no.1, pp.1-12.

3. Dhaliwal, J. and Wagner, J., 2021. A novel feature extraction method based on highly expressed SNPs for tissue-specific gene prediction. Journal of Big Data, 8(1), pp.1-13.

4. Clough, E., & Barrett, T. 2016. The Gene Expression Omnibus Database. In Statistical genomics, pp. 93-110.

5. JIAO, J., & LI, Y. 2017. Review of Typical Machine Learning Platforms for Big Data. Journal of Computer Applications, vol.37, no.11, p.3039.

6. Alatrany, A.S., Hussain, A.J., Mustafina, J. and Al-Jumeily, D., 2022. Machine Learning Approaches and Applications in Genome Wide Association study for Alzheimer's disease: A systematic review.pp. 62831-62847

7. Chandrashekar, G., & Sahin, F. 2014. A Survey on Feature Selection Methods. Computers & Electrical Engineering, vol.40, no.1, pp.16-28.

8. Srinivasan, B., & Pavya, K. 2017. Feature Selection Techniques in Data Mining: A Study. International Journal of Scientific Development and Research (IJSDR), vol.2, no.6, pp. 594-598.

9. Tadist, K., Najah, S., Nikolov, N.S., Mrabti, F. and Zahi, A., 2019. Feature Selection Methods and Genomic Big Data: A Systematic Review. Journal of Big Data, vol.6, no.1, pp.1-24.

10. Su, S.C., Lin, C.J. and Ting, C.K., 2011, December. An Effective Hybrid of Hill Climbing and Genetic Algorithm for 2D Triangular Protein Structure Prediction. In Proteome science, Vol. 9, No. 1, pp. 1-8. BioMed Central.

11. Uppu, S., Krishna, A. and Gopalan, R.P., 2016. A Deep Learning Approach To Detect SNP Interactions. J. Softw., vol.11, no.10, pp.965-975.

12. Ji, S., 2019. Deep Learning for Genome-Wide Association Studies and the Impact of SNP Locations (Doctoral dissertation, Memorial University of Newfoundland), pp. 1-109.

13. Batnyam, N., Gantulga, A., & Oh, S. 2013. An Efficient Classification for Single Nucleotide Polymorphism (SNP) Dataset. In Computer and information science, pp. 171-185. Springer, Heidelberg.

14. Alzubi, R., Ramzan, N., Alzoubi, H., & Amira, A. 2017. A Hybrid Feature Selection Method for Complex Diseases SNPs. IEEE Access, 6, pp. 1292-1301.

15. Shah, S. C., & Kusiak, A. 2004. Data Mining and Genetic Algorithm Based gene/SNP Selection. Artificial intelligence in medicine, 31(3), pp. 183-196.

16. Podder, M., Welch, W. J., Zamar, R. H., & Tebbutt, S. J. 2006. Dynamic Variable Selection in SNP Genotype Autocalling from APEX Microarray Data. BMCbioinformatics, vol. 7, no.1, pp. 1-11.

17. Silva, P. P., Gaudillo, J. D., Vilela, J. A., Roxas-Villanueva, R. M. L., Tiangco, B. J., Domingo, M. R., & Albia, J. R. 2022. A Machine Learning-Based SNP-Set Analysis Approach for Identifying Disease-Associated Susceptibility Loci. Scientific reports, vol.12, no.1, pp. 1-10.

18. Jiao, Y., Chen, R., Ke, X., Cheng, L., Chu, K., Lu, Z., & Herskovits, E. H. 2012. Single Nucleotide Polymorphisms Predict Symptom Severity of Autism Spectrum Disorder. Journal of autism and developmental disorders, vol.42, no.6, pp. 971-983.

19. Begum, A., & Parkavi, A. (2019). Prediction of Thyroid Disease Using Data Mining Techniques. In 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), pp. 342-345.

20. Nunes, C. M., Britto, A. S., Kaestner, C. A., & Sabourin, R. An optimized hill climbing algorithm for feature subset selection: Evaluation on handwritten character recognition. In Ninth International Workshop on Frontiers in Handwriting Recognition, pp. 365-370. IEEE.

21. Li, X. 2017. A Fast and Exhaustive Method for Heterogeneity and epistasis analysis based on Multi-Objective Optimization. Bioinformatics, vol.33, no.18, pp. 2829-2836.

22. Jing, P. J. & Shen, H. B. 2015MACOED: A Multi-Objective Ant Colony Optimization Algorithm for SNP epistasis Detection in Genome-Wide Association Studies. Bioinformatics 31, pp. 634–641.

23. Li, X., Liu, L., Zhou, J. and Wang, C., 2018. Heterogeneity Analysis and Diagnosis of Complex Diseases Based on Deep Learning Method. Scientific reports, vol. 8, no.1, pp.1-8.

24. Hess, M., Lenz, S., Blätte, T. J., Bullinger, L. & Binder, H.2017. Partitioned Learning of Deep Boltzmann Machines for SNP data. Bioinformatics 33, pp. 3173–3180.

25. Boutorh, A. and Guessoum, A., 2016. Complex diseases SNP Selection and Classification by Hybrid Association Rule Mining and Artificial Neural Network-Based Evolutionary Algorithms. Engineering Applications of Artificial Intelligence, 51, pp.58-70.

26. Nguyen, Q.T., Lee, E.J., Huang, M.G., Park, Y.I., Khullar, A. and Plodkowski, R.A., 2015. Diagnosis and Treatment of Patients with Thyroid Cancer. American health & drug benefits, vol.8, no.1, p.30.

27. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16619

28. Pirmoradi, S., Teshnehlab, M., Zarghami, N. and Sharifi, A., 2020. A Self-Organizing Deep Auto-Encoder Approach for Classification of Complex Diseases using SNP Genomics Data. Applied Soft Computing, 97, p.106718.

29. Urbanowicz, R.J., Meeker, M., La Cava, W., Olson, R.S. and Moore, J.H., 2018. Relief-Based Feature Selection: Introduction and Review. Journal of biomedical informatics, 85, pp.189-203.

30. Aldehim, G., De La Iglesia, B. and Wang, W., 2014, Heuristic Ensemble of Filters for Reliable Feature Selection. In ICPRAM, pp. 175-182.

31. Alzubi, R., Ramzan, N., Alzoubi, H. and Amira, A., 2017. A Hybrid Feature Selection Method for Complex Diseases SNPs. IEEE Access, 6, pp.1292-1301.