RESEARCH ARTICLE OPEN ACCESS

Scheming a Robust Diagnostic Signature for Cutaneous Melanoma with New Hybrid Classifier in Multimodal Data

V.Poonkodi [1], Dr. R.Hemalatha [2]
Research Scholar [1], HOD [2]
Department of Computer Science
Kumaran Atrs & science College for Women, Tirupur
Tamil Nadu - India

ABSTRACT

An integrated framework learning multi-modal datasets will target underlying causative biological actions that through a systems level disease demonstration are interpreted to the disease constitution. Intended by this, we aimed to associate low-level biological information to Cutaneous Melanoma (CM) disease status and compare the knowledge content of genes to it of macroscopic CM disease descriptors. Towards our goals, we tend to use two completely different datasets regarding CM. The datasets used here come back from two completely different sets of subjects that are represented either by gene expression or imaging options. To extend the potency of the system, we tend to apply completely different selection procedures to seek out the simplest set of options and different solutions of classifier. Exploitation applied math entropy-based ways at the side of useful analysis; we object a gene signature for CM, whereas imaging features designated once more statistically square measure compared to the chosen genes subset in terms of their information content. Elected genes were wont to train the new planned hybrid classifier known as Support Vector Machines with Cat Swarm optimization algorithm (mCSOA-SVM) that would generalize well once discriminating malignant from benign melanoma samples. The numerical results demonstrated that classifiers performed higher once designated genes were used as input, instead of imaging features designated by information gain measurements. Regarding the accuracy of the planned recognition system have confirmed sensible accuracy of the projected methodology and high sensitivity in malignant melanoma recognition.

Keywords:- Multimodal Data, Classification, Composite Biomarkers, Cutaneous Melanoma, Dermoscopy, Feature Selection, Gene Ontology, Image Analysis, Microarrays.

I. INTRODUCTION

Cutaneous melanoma may be a disease of accelerating clinical and economic importance within the world. The danger of death from malignant melanoma is expounded on to the Breslow thickness of the first lesion. This correlation has been documented thus well that Breslow thickness at presentation are often wont to predict mortality from the disease confidently. This reality makes efforts at secondary interference through early detection notably vital. There are solely few reports within the literature on CMM within the Asian population. The Regional Cancer Centre at Trivandrum registers concerning 6000 new cancer cases annually and CMM forms 0.5% of them. This paper presents the clinical medicine and survival of seventy nine cases of CMM registered within the hospital cancer written record throughout the amount 1985-90. Since there is no effective medical care for patients with advanced malignant melanoma, instructional campaigns decide to encourage insecure people to bear routine screening so melanomas are often known early whereas they are still simply treatable. These academic campaigns generate an oversized quantity of referrals to dermatologists whom services are already underprovided.

Integration of multimodal and multiscale knowledge is of known importance within the context of customized drugs and future electronic health record management. The search for appropriate knowledge fusion schemes, that may ideally optimize the exploitation of the knowledge residing in composite datasets, is an aborning space with varied potential applications.

Within the context of Virtual Physiological Human (VPH), associate degree integrated framework promotes the interconnection of predictive models pervading totally different scales, with totally different ways, characterized by completely different coarseness. Such a framework consolidates system level data and permits formulation and testing of hypotheses, facilitating a holistic approach [1].

In this work we tend to propose a unique methodology on multimodal knowledge fusion relating to separate datasets. As separate we outline datasets wherever every have been obtained from a unique technological supply and from a unique set of patients. Even the quantity of patients taking part in every examination is not an equivalent. The sole common determinant of separate datasets is that they discuss with an equivalent disease. Such datasets don't seem to be amenable to normal fusion ways, as all known ways trot out an equivalent set of patients being examined by varied instruments and techniques in sequence, so manufacturing the multimodal knowledge. However, the bulk of open-accessible knowledge refers to the unimodal results of bound experiment relative to a particular disease. The advised methodology is ready to focus on biomarkers utilizing these separate unimodal outcomes and so repurpose the prevailing knowledge of accessible repositories.

As proof of thought, we have a tendency to specialize in the fusion of two separate unimodal datasets, one amongst molecular and one amongst imaging description, each involved with the study of CM disease. Purpose of feature selection and spatiality reduction algorithms on the created unified dataset will contribute towards the extraction of higher biomarkers, ruling out false positive findings synchronous, however with no causative association, with the investigated disease. The procedures are often applied to numerous cases and tackle separated datasets of alternative diseases still. The chosen descriptors are used because the input attributes to the system of classifiers (Support Vector Machines with Cat Swarm optimisation Algorithm), answerable for the ultimate recognition of malignant melanoma.

This paper is organized as follows: Section two includes connected work and background on data fusion ways, the cutaneous melanoma disease, and also the feature choice techniques utilized in this work. Section 3 describes the preprocessing steps for the preparation of the multimodel dataset and feature selection with proposed classifier. Section four encloses the results of the feature choice procedures relating to specific biomarkers and their performance and stability determined during repetitive runs. Finally, Section five concludes with future work.

II. RELATED WORK

Cutaneous melanoma (CM): It is considered a complex multigenic and complex disease that involves each environmental and genetic factor. It is the foremost grievous tumour of the skin, and its incidence and mortality are perpetually increasing worldwide. CM tumorigenesis is usually explained as a progressive transformation of traditional melanocytes to nevi that later turn into primary cutaneous melanomas (PCM). However, the molecular pathways concerned haven't been clearly elucidated, though significant progress has been created [2]. Despite the success of genetics in process genomic markers or gene signatures for other forms of cancers (such as breast cancer), there has been no similar progress involving melanoma.

The microarray studies that are performed on CM by totally different teams exploit different microarray technological platforms applied in extremely heterogeneous patient cohorts and pathological sample collections [3]. These variations hurdle considerably comparisons, yielding cohorts of reduced total size and variety. Integration of freelance cohorts from totally different studies bears vital challenges for variety of reasons stemming from the technical style to strictly biological ones [4].

Regarding the clinical ways for designation of malignant melanoma, there exist many customary approaches for analysis and designation of lesions, let's say, the Menzies, scale, the Seven-point scale, the overall Dermoscopy Score supported the ABCD rule, and also the ABCDE rule (Asymmetry, Border, Color, Diameter, Evolution). In these ways, digital images will be a basis for the medical analysis and designation of lesions into account. As human interpretation of image content is fraught with discourse ambiguities, advanced processed techniques will assist doctors

within the diagnostic method [5]. A review of image acquisition and have extraction ways utilized within the literature concerning existing classification systems will be found in [6].

Feature Selection: concerning the applied feature selection procedures during this study, at first, a wrapper kind technique was used (sequential backward elimination—SBE) mistreatment the random forest (RF) algorithm [7], which utilizes ensembles of call trees. South by east algorithm starts with the total set of options and iteratively removes the feature computed as least necessary on every occasion, till a needed variety of options stay. As an option, a variable filter was wont to cut back the colinearity among options of the microarray dataset, before the applying of the wrapper methodology. This filtering at the side of the imputation represents a transition from a COD methodology towards a GFF approach, though here no additional transformation is applied to the feature vectors.

The random forest algorithm, among different ensemble learning ways, is rumored to achieve success in variance reduction, which is related to reducing overfitting [8]. Additionally, we tend to use the choice of stratifying the bootstrapped samples with equal variety of cases per category [9]. This is often compatible with the Balanced Random Forest (BRF) approach, that is computationally additional economical with massive unbalanced knowledge, since every tree solely uses alittle portion of the training set to grow. To boot it's less at risk of noise (mislabelled category) than the Weighted Random Forest (WRF) wherever a heavier penalty is placed on misclassifications of the minority class [10].

PCA defines new orthonormal intermediate variables, consisting of linear mixtures of the initial variables specified the axis of the first principal component (PC) highlights the dimension linear to the foremost variation and also the axis of the second part the dimension with the foremost of the remaining variation so on. The coordinates of the samples within the new house created by the PCs are known as scores [11].

LDA uses category info to maximise the separation between numerous teams of observations. LDA presumes that the categoryification variables follow a traditional variable distribution and also the variance matrices for the observations of every class are equal (homoscedasticity). Once these operating assumptions aren't thought-about plausible, LDA doesn't represent the optimum classifier. However, it will still be thought-about a legitimate and correct methodology for the screen of the dimensional resolution house, once the target is that the determination of separating hyperplanes that maximize discrimination between totally different categories. This is so; as a result of the hypotheses on the shape of the info distributions don't have any impact on the answer of the geometrical separation downside [12].

III. MATERIALS AND METHODS

3.1 Multimodal Data

The progress of the methodology is shown in Figure 1. At the initial section (a) there are two unimodal separate datasets (image and microarray data). Every table has totally different features (If and Mf) and different numbers of observations/rows (Ir and Mr) obtained from different patients. The last column of the tables represents the response variable (rv). In our case it is a binary response variable with two classes: healthy or malady. Next (b) the unimodal tables are incorporate to at least one block thin matrix. The sole column while not nonavailable values is that the one with the response variable. Afterwards, at step (c) straightforward biased imputations are performed per feature and per category. These are delineating with dotted lines. The entire range of rows and options of the unified table is the sum of the rows and options of the 2 initial tables (Ur, Uf). Currently the table is amenable to variable statistical analysis, and specific composite biomarkers will be extracted and studied for their performance contribution and stability in their look over repetitive runs of synthetic data creation via the imputations.

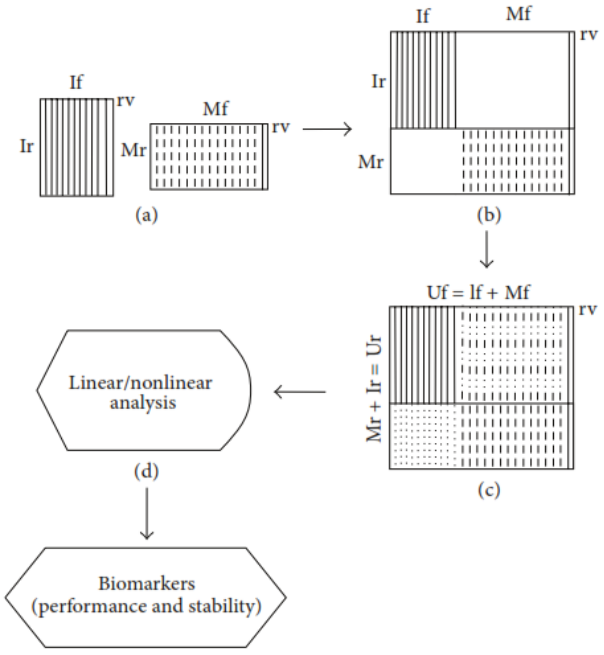

Figure 1: data fusion advancement for separate datasets: (a) separate datasets, (b) unified thin dataset, (c) unified dataset (class imputations), and (d) and (e) variable statistical analysis and feature selection.

Image data

The dataset derived from skin lesion pictures contained 972 instances of defect skin lesions and sixty nine skin cancer cases. Three forms of features were analyzed: border features that cover the a and b components of the abcd-rule of medicine, color features that correspond to the c rules, and textural features that are supported d rules. Thirty one out of the initial set of thirty two doable features were used; one feature was removed because of having zero variation across the samples. The relevant preprocessing for all features is delineate in [13]. The scale of the image dataset were so 1041 (rows) thirty one (columns).

Microarray data

The microarray dataset was taken from the gene expression omnibus (geo) [14], gds1375. In this experiment, total polymer isolated from forty five primary skin cancer, eighteen benign skin nevi, and seven traditional skin tissue specimens was used for gene expression analysis, mistreatment the affymetrix hu133a microarray chip containing 22,000 probe sets [15]. The dataset contains the mas5-normalized signal intensities and is globally scaled in order that the typical intensity equals 600.

Data retrieval from geo was performed mistreatment geoquery [16] and concomitantly processed with limma [17] r packages from the bioconductor project [18], following the planned steps as listed within the r script made by the geo2r tool [19]. The gene expression values across all classes were log-transformed, and therefore the mean values of all genes within the traditional skin were calculated. Afterwards, the mean sequence vector regarding the traditional skin classes was ablated from all replicate vectors of the other two classes. During this manner, the initial signal intensities provided ratios of differential expression, calculated by dividing the signal intensities of every class by the individual sequence worth of the traditional class. As all values are log-transformed, the division is replaced with a subtraction. For the remaining analysis the differentially expressed sequence values of the melanoma versus skin and nevi versus skin were exploited. 1701 genes from a linear model match were extracted setting fdr for multiple testing adjustment, value 0.001 and 2-fold changes as thresholds. The scale of the microarray dataset were so 63 (rows) 1701 (columns).

Data integration

The two tables containing the microarray and image data were incorporate to at least one block thin matrix with dimensions 1104 rows 1734 columns, marking the inaccessible values as not available (na). The rows contain the microarray and image information samples and therefore the columns microarray and image features and one binary response variable (0 for defect and 1 for melanoma).

Missing values imputation

Although there are many software packages implementing advanced imputation strategies [20], they may not be used during this unified informationset wherever the multimodal data have solely the category variable column as complete. During this study we have a tendency to thought of four easy imputation strategies applied per feature and per class:(i)“mean value” imputation,(ii)“random

normal" imputation,(iii)"uniform" imputation,(iv)"bootstrap" imputation.

In the second case, once estimating the average (m) and variance (sd) of every feature (ignoring the sodium values) per category, we tend to every which way stuffed the missing values sampling from associate assumed distribution having as parameters: (m, sd). The "uniform" imputation is conducted by sampling uniformly at intervals vary of every feature per category and therefore the "bootstrap" imputation by independent bootstrap of every variable singly per category, till all the sodium values are replaced. The last 2 imputation strategies are like the manner random forests construct synthetic data, so as to supply for a similarity measure. For the economical execution of the imputations, the plyr R package was utilized [21].

3.2 Feature Selection with Functional Analysis and Information Gain

The GOwas employed in order to explore the underlying useful content at a lower place the set of 1701 genes differentially expressed between the benign and melanoma samples. To the present finish, the GORevenge algorithm developed by the authors [22] was used. GORevenge exploits graph-theoretical algorithmic methodologies and consistently exploits the GO tree so as to help the elucidation of hidden useful regulative effects among genes. As a way of inferring potential useful cliques among genes, GORevenge uses a stepwise mechanism, ranging from the AB initio thought of gene set. Within the first (agglomeration) section, genes are collected given their links to a given GO term, and to its neighboring ones, further its folks and kids GO terms. These genes are thought of to belong to an equivalent functional clique, which is defined by the employment of distance-based useful similarity criteria. Specifically, genes were hierarchical supported their IG ratio values [23], measured because the average of IG ratios altogether six datasets (three uniform and 3 bootstrap) derived by the 2 data imputation strategies. Genes with an IG ratio worth within the high two hundredth (340 genes) were designated because the most informative ones in terms of variability. Genes with comparatively high IG and genes with a central regulative role within the underlying active molecular networks (after applying the GORevenge algorithm) represent genes with reciprocally independent characteristics. Therefore, an intersection of the high IG valued gene set with the functionally connected GORevenge sequence sets may reveal the extremely informative crucial molecular players concerned within the development of CM. the two GORevenge output lists (MF, BP) were, thus, intersected with genes that conferred high immune globulin and thirty two genes found common comprised the gene signature conferred here. The set of thirty one imaging options was conjointly evaluated in terms of immune globulin magnitude relation in relevance malady standing (malignant versus benign) and were prioritized per immune globulin ratios obtained.

3.3 Further Selection (Genes/Imaging) by Total Variance of Feature Subsets

Prioritizing Imaging features From the subsets of thirty two designated genes and thirty one imaging features, the correlate options were removed (correlation coefficient>0.8) so as to get rid of redundant features. No redundancies was found within the genes set, whereas a number of the imaging features (8 out of the 31) were removed as were found correlate with different imaging features (See Section IV). The hierarchical designated genes and imaging features per immune globulin ratios in relevance malady standing were employed in order to gather a set of features (genes or imaging) that carries an adequate total variance (TV) to the malady standing. Thus, the TV was calculated for progressive subsets, the first one adore the first hierarchical feature and therefore the last one to the complete set of designated features (post all previous feature choice applied). TV was calculated because the total of normalized ig ratios (IG ratio/sum of ig ratios of all features) for the features enclosed in every of the progressive feature subsets fashioned.

For each of the feature subsets, four classifiers were made and evaluated in terms of generalization using singly all six informationsets (three uniform and 3 bootstrap) derived by the 2 data imputation strategies used here. Specifically, k-nearest neighbor (k-NN) classifiers [24], a call tree (DT) [25], the random forest (RF) algorithm [26] and therefore the planned hybrid classifier were used. Their performance was measured

mistreatment three-cross validation resampling. Sensitivity measurements (true positive/(true positive + false negative) for every of the two categories all told six datasets and mean values were calculated and accustomed compare the chosen genes set and imaging features set.

3.4 Proposed Hybrid mCSOA-SVM Classifier

Neither seeking nor tracing mode is capable of retentive the most effective cats; but, activity a local search close to the most effective answer sets will facilitate within the choice of higher answer sets (selection of best featureset). This paper proposes a modified cat swarm optimisation Algorithm (mcsoa) to boost looking out ability within the neighborhood of the simplest cats. Before examining the method of the algorithmic rule, we have a tendency to should examine variety of relevant problems. Classifiers area unit algorithms wont to train information for the development of models won't to assign unknown information to the classes during which they belong. This work adopted an SVM as a classifier. SVM theory comes from statistical learning theory, supported structural risk step-down. Svms are wont to notice a hyperplane for the separation of two teams of information. This work regards the SVM as a recorder that receives coaching information for the classification model.

Solution Set Design

n answer set containing information in two parts: SVM parameters (C andγ) and have subset. Parameter C is that the penalty constant which means tolerance for error. And, parameter γ may be a kernel parameter of RBF, which implies radius of RBF. The continual worth of those two parameters is regenerate into binary writing. Another is feature subset variables ($F_1 \sim F_n$) wherever n is that the range of options. The change of the variables in every feature subset falls between zero and one. If F_i is bigger than zero.5, its corresponding feature is selected; otherwise, the corresponding feature is not chosen.

C	γ	F_1	$\cdots$ F_i $\cdots$	F_n

Figure 2: Representation of a solution set.

Mutation Operation

Mutation operators are wont to find new answer sets within the neighborhood of different answer sets. Once an answer set mutates, each feature encompasses likelihood to vary. Additionally, C and γ should be modified to binary, in order that they will choose a mutation operation for the subsequent search.

Evaluating Fitness

This study utilized k-fold cross-validation [24] to check the search ability of the algorithms. The k was set to five, indicating that 80% of the initial information was every which way designated as coaching information and therefore the remainder was used as testing information. And so retain features up to an answer set that we wish to judge for coaching information and testing information. This coaching information is input into an SVM to make a classification model used for the prediction of testing information. The prediction accuracy of this model represents its fitness. So as to match answer sets with identical fitness, we tend to thought of the amount of designated options. If prediction accuracy were identical, the answer set with fewer designated features would be considered superior.

Algorithm for planned mCSOA-SVM

The steps of the planned mCSOA-SVM are conferred as follows.

Step 1. Every which way generate N solution sets and velocities with D-dimensional house, depicted as cats. outline the subsequent parameters: seeking memory pool (SMP), seeking range of the selected dimension (SRD), counts of dimension to change (CDC), mixture ratio (MR), number of best solution sets (NBS), mutation rate for best solution sets (MR_Best), and range of attempting mutation (NTM).

Step 2. Judge the fitness of each solution set using SVM.

Step 3. Copy the NBS best cats into best solution set (BSS).

Step 4. Assign cats to seeking mode or tracing mode supported mister.

Step 5. Perform search operations adore the mode (seeking/tracing) allotted to every cat.

Step 6. Update the BSS. For each cat once the looking out method, if it's higher than the worst solution set in BSS, then replace the worst solution set with the higher solution set.

Step 7. For every solution set in BSS, search by mutation operation for NTM times. If it's higher than the worst answer set in BBS, then replace the more severe solution set with the higher solution set.

Step 8. If terminal criteria area unit glad, output the simplest subset; otherwise, come back to (step 4).

After looking out in accordance with seeking and tracing modes, the cats with higher fitness will be updated to the simplest solution set (BSS). A mutation operation is then applied to BSS to go looking different solution sets. If a solution set following mutation shows improvement, it replaces the initial solution set within the BSS.

IV. RESULTS AND DISCUSSION

GORevenge analyses are resulted in total of 179 genes found within the original list of 1701 differentially expressed genes. Out of this set of genes, thirty two genes had an IG ratio within the high 20% of all 1701 genes, which are characterized by a high info content measured with entropy-based statistics (mean IG ratio > 0.81) and are coupled to a central restrictive role as derived by the GOrevenge analysis. These genes comprise a gene signature underlying manifestation, on the idea of the accessible gene expression profiling information. The correlation correlation coefficient is greater than 0.8 were found within the gene set. The TV every of the progressive genes set, ranging from the first graded gene supported IG ratio up to any or all thirty two genes, was calculated.

TV measurements show that a gene set comprising of twenty six genes (MAP3K4, CTSB, ...,PBX1) carries most of the variation content to the disease standing and corresponds to measure of TV > zero.8. The use of this total variation criterion is accustomed more slender down the spatial property of the set of elect genes set. The imaging features were found to be less related to illness standing (malignant versus benign) than genes in step with Ig ratios. The highest 10 imaging features with an Ig ratio > 0.15 were elect because the ones that comprise an imaging feature set derived based mostly only on statistics, i.e., the data content to disease standing supported IG ratios. Correlate imaging features where correlation coefficient>0.8) and twenty three imaging featured were unbroken.

The TV every of the progressive imaging features set carries, ranging from the first graded gene supported IG quantitative relation up to any or all twenty three unrelated features, was calculated. TV measurements show that a feature set comprising of thirteen features (mean.R, Area,.., Distance.std) carries most of the variation content to the illness standing and corresponds to measure of TV > 0.8. The use of this total variation criterion was accustomed more slender down the spatial property of the set of elect imaging features. We tend to next used the twenty six elect genes (TV > 0.8) and also the thirteen imaging features (uncorrelated imaging features, TV > 0.8) to prioritise the imaging features in step with their correlation to the genes within the shortest list of important molecular players.

This was done by ranking the imaging features per gene in step with MI values to organic phenomenon values. A careful observation of the imaging features are sorted at the terribly prime for all or most genes such as mean.R, mean.G. Specifically, the highest 10 imaging features in step with the prioritization overall all twenty six elect genes correspond to mean.R, mean.G, std.R, mean.B, S.std, S.mean, unsimilarity, complexity, Grad.mean, and spatial property. Their MI measurements to the chosen genes were found to be within varying (0.056–0.148) showing moderate, nevertheless existing, mutual info with the genes. These 10 imaging features comprise another set of features which will be evaluated for the discrimination between malignant and benign cases.

The GA-based methodology used for imaging features selection was set to yield the most effective activity ten-dimensional feature set, out of the first set of thirty one imaging features, in terms of the accuracy obtained by the hybrid classifier as represented within the previous section. Once GA convolution, a set as well as the features of mean.R, GMSM.mean, S.mean, L.mean, A.mean, B.mean, B.std, Grad.mean, Grad.std,

and Grad.max. In total four gene or imaging features subsets were derived by the methodologies antecedently described: 1) thirty-two factors comprising the gene signature (functional analysis and IG ratios), 2) 10 imaging options elect based mostly only on info content to disease status (IG ratios), 3) 10 imaging options elect supported the prioritization to the list of genes carrying a TV >; 0.8 by means of MI values, and iv) 10 imaging options elect by the GA-based choice theme. All four set of features were input to the classifiers used here. Sensitivity results for the two categories (malignant, benign) obtained for all six instances of the integrated dataset and mean values, are given for these four features subsets.

3.5 False positive rate (FPR) Comparison

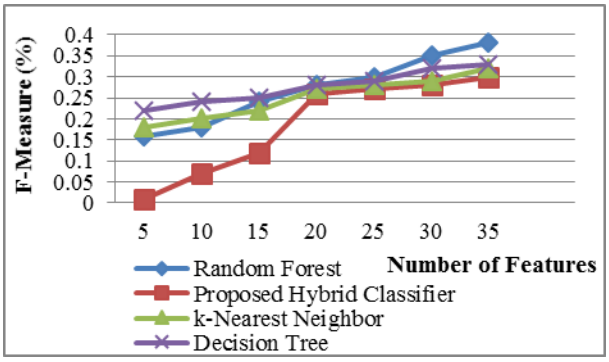

Fig.3. False Positive Rate Comparison

The graph of false positive rate in percentage is shown in Figure.3. The proposed method have low false positive rate 10% which is significantly enhance detection rate and accuracy. Other method like RF has 16%, k-NN has 18% and DT have 22% detection rate.

3.6 Accuracy Comparison

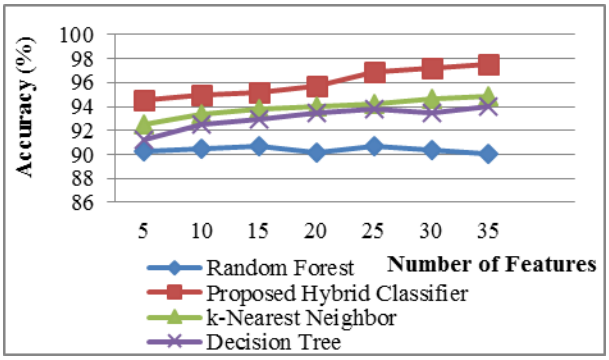

Fig.4. Prediction and Detection accuracy Comparison

Prediction and detection accuracy of the proposed method achieves higher classification accuracy than the existing classification methods such as RF, DT and k-NN, and the accuracy result is illustrated in Figure 4 as the proposed methods selects the important features in the feature subset. In proposed work perform CSO in order to reduce the errors.

3.7 Execution Time comparison

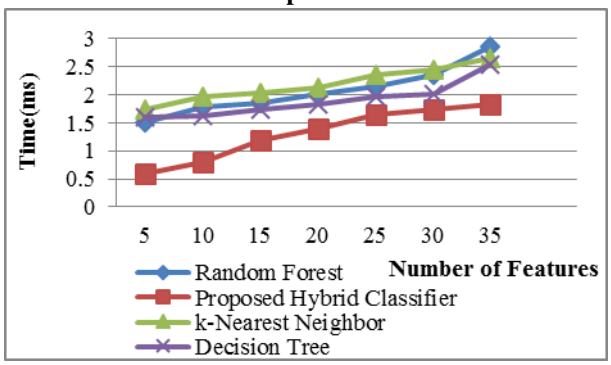

Fig.5.Execution Time comparison

Figure 5 shows the execution time comparison results of the proposed approach and the existing methods such as such as RF, DT and k-NN. The proposed approach has less execution time to detect the CM disease, since the proposed method can effectively predict the possibility of disease features through an effective CSO.

V. CONCLUSION AND FUTURE WORK

In this work, an integrated dataset concerning CM was accustomed derive an inventory of thirty two genes extremely concerning CM disease status that comprise potential CM biomarkers. Statistics and a biologically impressed selection scheme that measures the impact of genes to imaging features were accustomed prioritize a complete of thirty one imaging features and choose the foremost strong ones. Victimization appropriate classification techniques, imaging features may discriminate with a moderate success malignant from benign CM samples.

During this classification task used a brand new hybrid classifier, imaging features were outperformed by the genes set elect here because the latter contained abundant denser info on the manifestation of CM. Still, this approach may well be more developed to implement automatic feature selection, exploiting statistically derived call cut-offs. It also can be extended to alter juxtaposition with the machine-driven filtering that happens within the feature selection with RF. Additionally, the signatures derived by the unified datasets may well be compared with clinical multimodal data stemming from constant set of patients. During this means, the impact of the imputation strategies within the creation of the synthetic dataset may well be cross-evaluated, with the performance of different information analysis strategies, corresponding to canonical correlation analysis or variations of such strategies.

REFERENCES

[1] J. W. Fenner, B. Brook, G. Clapworthy et al., "The EuroPhysiome, STEP and a roadmap for the virtual physiological human," Philosophical Transactions of the Royal Society A, vol. 366, no. 1878, pp. 2979–2999, 2008.

[2] M. Balázs, S. Ecsedi, L. Vízkeleti, and A. Bégány, "Genomics of human malignant melanoma," inBreakthroughs in Melanoma Research, Y. Tanaka, Ed., InTech, 2011.

[3] J. Tímár, B. Gyorffy, and E. Rásó, "Gene signature of the metastatic potential of cutaneous melanoma: too much for too little?" Clinical and Experimental Metastasis, vol. 27, no. 6, pp. 371–387, 2010.

[4] W. K. Martins, G. H. Esteves, O. M. Almeida et al., "Gene network analyses point to the importance of human tissue kallikreins in melanoma progression," BMC Medical Genomics, vol. 4, article 76, 2011.

[5] M. Ogorzałek, L. Nowak, G. Surówka, and A. Alekseenko, "Modern techniques for computer-aided melanoma diagnosis," in Melanoma in the Clinic—Diagnosis, Management and Complications of Malignancy, M. Murph, Ed., InTech, 2011.

[6] I. Maglogiannis and C. N. Doukas, "Overview of advanced computer vision systems for skin lesions characterization," IEEE Transactions on Information Technology in Biomedicine, vol. 13, no. 5, pp. 721–733, 2009.

[7] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[8] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: a review," International Journal of Pattern Recognition and Artificial Intelligence, vol. 23, no. 4, pp. 687–719, 2009.

[9] A. Liaw and M. Wiener, "Classification and regression by randomForest," R News, vol. 2, no. 3, pp. 18–22, 2002.

[10] C. Chen, A. Liaw, and L. Breiman, "Using random forest to learn imbalanced data," 2004, http://stat-reports.lib.berkeley.edu/accessPages/666.html.

[11] W. N. Venables and B. D. Ripley, Modern Applied Statistics with S, Springer, New York, NY, USA, 2002.

[12] Z. Díaz, M. J. Segovia, J. Fernández, and E. M. d. Pozo, "Machine learning and statistical techniques. An application to the prediction of insolvency in spanish non-life insurance companies," The International Journal of Digital Accounting Research, vol. 5, no. 6, pp. 1–45, 2005.

[13] M. Maragoudakis and I. Maglogiannis, "Skin lesion diagnosis from images using novel ensemble classification techniques," in Proceedings of the 10th IEEE EMBS International Conference on Information Technology Applications in Biomedicine, Corfu, Greece, November 2010.

[14] T. Barrett, D. B. Troup, S. E. Wilhite et al., "NCBI GEO: archive for functional genomics data sets-10 years on," Nucleic Acids Research, vol. 39, no. 1, pp. D1005–D1010, 2011.

[15] D. Talantov, A. Mazumder, J. X. Yu et al., "Novel genes associated with malignant melanoma but not benign melanocytic lesions," Clinical Cancer Research, vol. 11, no. 20, pp. 7234–7242, 2005.

[16] D. Sean and P. S. Meltzer, "GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor," Bioinformatics, vol. 23, no. 14, pp. 1846–1847, 2007.
[17] G. K. Smyth, "Limma: linear models for microarray data," in Bioinformatics and Computational Biology Solutions Using R and Bioconductor, pp. 397–420, Springer, New York, NY, USA, 2005.
[18] R. C. Gentleman, V. J. Carey, D. M. Bates et al., "Bioconductor: open software development for computational biology and bioinformatics," Genome Biology, vol. 5, no. 10, p. R80, 2004.
[19] NCBI_GEO, "GEO2R," http://www.ncbi.nlm.nih.gov/geo/info/geo2r.html.
[20] N. J. Horton and K. P. Kleinman, "Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models," American Statistician, vol. 61, no. 1, pp. 79–90, 2007.
[21] H. Wickham, "The split-apply-combine strategy for data analysis," Journal of Statistical Software, vol. 40, no. 1, pp. 1–29, 2011.
[22] K. Moutselos, I. Maglogiannis, and A. Chatziioannou, "GOrevenge: A novel generic reverse engineering method for the identification of critical molecular players, through the use of ontologies," IEEE Trans. Biomed. Eng., vol. 58, no. 12, pp. 3522–3527, Dec. 2011.
[23] D. Jianhua and X. Qing, "Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification," Appl. Soft Comput., vol. 13, no. 1, pp. 211–221, 2013.
[24] Y. Wu, K. Ianakiev, and V. Govindaraju, "Improved k-nearest neighbor classification pattern recognition," Comput. Vision Pattern Recog.. vol. 35, pp. 2311–2318, 2002.
[25] L. Breiman, J. H. Friedman, R. A. Olson, and C. J. Stone, Classification and Regression Trees. Belmont, CA, USA: Wadsworth, 1984.
[26] L. Breiman, "Random Forests," Mach. Learning, vol. 45, no. 1, pp. 5–32, 2001.