Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 221–226

ISBN 978-83-60810-51-4

Improved Feature Selection for Hematopoietic Cell Transplantation Outcome Prediction using Rank Aggregation

Chandrima Sarkar
College of Science and Engineering
University of Minnesota at Twin Cities
Minneapolis, MN 55455

Sarah Cooley
Hematology, Oncology and Transplantation
University of Minnesota at Twin Cities
Minneapolis, MN 55455

Jaideep Srivastava
College of Science and Engineering
University of Minnesota at Twin Cities
Minneapolis, MN 55455

***Abstract*—This paper presents a methodology for developing an improved feature selection technique that will help in accurate prediction of outcomes after hematopoietic stem cell transplantation (HSCT) for patients with acute myelogenous leukaemia (AML). Allogeneic HSCT using related or unrelated donors is the standard treatment for many patients with blood related malignancies who are unlikely to be cured by chemotherapy alone, but survival is limited by treatment-related mortality and relapse. Various genetic factors such as tissue type or human leukocyte antigen (HLA) type and immune cell receptors, including the killer-cell immunoglobulin-like receptor (KIR) family can affect the success or failure of HSCT. In this paper we aim to develop a novel, aggregated ranking based feature selection technique using HLA and KIR genotype data, which can efficiently assist in donor selection before BMT and confer significant survival benefit to the patients. In our approach we use a rank aggregation based feature selection technique for selecting suitable donor genotype characteristics. The result obtained is evaluated with classifiers for prediction accuracy. On average, our algorithm improves the prediction accuracy of the results by 3-4% compared to generic analysis without using feature selection or single feature selections algorithms. Most importantly the selected features completely agree with those obtained using traditional statistical approaches, proving the efficiency and robustness of our technique which has great potential in the medical domain.**

I. Introduction

APPROXIMATELY 12,000 cases of Acute Myelogenous Leukaemia (AML) are diagnosed annually in the United States. Many patients are not cured by chemotherapy alone, and require hematopoietic stem cell transplantation (HSCT) for curative therapy. While HSCT can cure AML, it is a complex procedure with many factors influencing the outcomes, which remain suboptimal [1]. Donor selection is a critical part of the entire transplant procedure and researchers are looking for host or donor genetic factors that can predict a successful outcome after transplantation. For allogeneic HSCT to be successful, the leukemia cells must be eradicated by the, combined effect of chemotherapy, radiotherapy, and a donor T cell mediated graft-versus-leukemia reaction. The donor stem cells reconstitute the patients ablated hematopoietic and immune systems which is important to prevent relapse and prevent infections[2], [3]. The most important factor in donor selection is matching for human leukocyte antigens (HLA). In addition, other factors such as donor age, gender, parity, and prior exposure to viruses such as cytomegalovirusare considered as they can influence transplant outcomes[3]. Recently, investigators have focused on the role of natural killer (NK) cells on mediating beneficial effects in HSCT[4], [3] NK cells express polymorphic killer-cell immunoglobulin-like receptors (KIR)[4], [5] which influence the function of NK cells which can kill leukemia cells, decrease rates of graft versus host disease, and control infections after HSCT. Because the HLA genes and KIR genes are on separate chromosomes only 25% of HLA-matched sibling donors are KIR identical and unrelated HLA-matched donors are rarely KIR identical[6].In two papers analysing a retrospective cohort of patients receiving unrelated donor transplants for AML demonstrated the beneficial effect of certain donor KIR genes on preventing relapse and improving survival after HSCT [3]. The most important result was the identification of groups of KIR genes from the centromeric and telomeric portions of the genetic region which were associated with relapse protection and survival. Specifically, donors with KIR B haplotype genes were protctive. We chose this dataset to test the ability of this novel data mining based approach to identify relevant variables because of the complexity of HSCT and the high dimensional data with a large number of donor and recipient attributes. To the best of our knowledge, this approach has never been used in this domain. Machine learning techniques have never been explored to find patterns in genetic data to improve donor selection algorithms and predict outcome after HSCT.

To summarize, this paper makes the following contributions:

- Development of a novel ensemble feature selection technique designed to find the best donor match for patients undergoing a HSCT. Importantly, our approach gave an overall high prediction accuracy across a variety of classifiers using genetic and other clinical data.
- Accurate prediction of treatment related mortality, relapse and disease free survival rates for patients with AML using our features selection approach.
- The results of our work show that our feature selection algorithm can be used efficiently for high accuracy pre-

diction models. This research supports the conclusion that data mining can enhance analysis of data rich domains like medicine, where patients may benefit from detection of information hidden in the data.

The remainder of the paper is organized as follows. Section II is the Motivation and Related works. Section III is the Proposed approach section. Section IV describes the experimental results. Section V is the Conclusion and future works.

II. Motivation and Related Works

Feature selection techniques have immense potential to enhance data mining in the medical domain as has been previously studied in areas such as medical image processing[7], [8], [9], [10]. Ensemble feature selection techniques have been used in the past to improve robustness and accuracy, but little is known to have been done in the medical domain. Ensemble methods are advantageous because these can outperform the single feature selection models when weak or unstable models are combined, mainly because in many cases several different but equally optimal hypotheses may exist and the ensemble reduces the risk of choosing a wrong hypothesis. Another advantage of ensemble methods is that in contrast to learning algorithms, which may end up in different local optima, ensemble may give a better approximation of the true function[11], [12].

We chose this dataset in part because it is high dimensional with missing data, characteristic of real biologic data, and because it has been extensively studied by traditional biostastical methods to provide good gold standard results to compare to our findings. This dataset is unique in that the donor and recipients of URD HSCT were genotyped not only for their HLA alleles, but also for the NK receptor KIR genes. It is known that the interactions between KIR and HLA molecules (their natural ligands) affect the function of NK cells and their ablity to kill cancer cells and to function to fight infection and promote overall immunity[4], [5], [13], [14], [15]. Several studies have documented the interaction between HLA and KIR on outcomes after HSCT [16], [17], [18], [19]. The data set used here was described in the first study to demonstrate that both centromeric and telomeric KIR genes from group B haplotypes contribute to relapse protection and improved survival after URD HSCT for AML [3], [2].The authors performed multivariate statistical analyses to identify genetic factors related to KIR that improve outcome after HSCT. The models included many donor and recipient transplant and demographic variables known to affect the outcome of HSCT.

The previously published analyses of this data set were designed to develop a decision strategy to efficiently select the optimal donor to prevent relapse after transplant and to improve survival. The methodologies used in these studies were generally classical statistical tests of hypotheses generated by physicians trying to interpret a plethora of variables based on prior knowledge. However, this approach, while highly accurate, is time consuming and potentially limited by the biases of the researchers generating the hypotheses. In any medical condition, treatment decisions can be challenging. The ultimate decision especially in case of transplants, rest with the physicians, who may be overwhelmed with a confusing range of information sources. The data is huge in medical domain and human beings have a limited ability to retain information as compared to the artificial intelligence, and this worsens when the amount of information increases. As a result, often there may be undue influence from personal experience. In such situations data mining can be a blessing where the automated techniques of significant variable selection can provide to medical experts the advantage of having a supporting second opinion for a more accurate decision making. Using data mining, interesting rules and relationships can be sought and discovered without prior knowledge. Data mining in general helps to capture cumulative experience of all the patients reflected in the entire database which can exhibit unknown pattern of medical significance. In this regard, feature selection can prove to be a highly efficient approach for detecting the contributing variables from an entire database. The result obtained from feature selection is a set of highly significant variables which can be used for accurate prediction purpose, either using classification techniques or statistical approaches. In this paper we aim at providing the medical domain with a novel feature selection approach which will help the domain experts in donor selection for a successful HSCT outcome. In medical domains like oncology, no research is known to have been conducted to the best of our knowledge, using a efficient feature selection approach which can be utilized for successful prediction outcomes. Our research can be considered as the first known work in the development of an automated approach for features or variables selection towards developing a donor selection strategy for HSCT based on information obtained from a large clinical genotype data repository.

III. Proposed Approach

A. Preliminaries

The main methodology used in this research is a rank aggregation based feature selection technique on high dimensional genetic data, followed by classification of the data corresponding to the extracted features to verify the prediction accuracy. The rationale behind using feature selection is two fold. Firstly, to eradicate redundant features with minimum effect on the predicted outcomes and secondly, to capture features which may prove as essential factors during donor selection for a successful outcome. The novelty of our approach is the use of rank aggregation measure for feature selection. Our algorithm uses rank aggregation technique for feature extraction. The result obtained from the above is a list of significant and globally selected set of variables that can be used as a selection criteria when selecting donors for AML patients. The general implication of global ranking is that it helps to rule out biases caused by individual algorithms while providing higher accuracy, sensitivity, and specificity, which are often not achievable with single models or while not using any feature selection model at all[20].

In the final step, the feature set obtained was used for the prediction of survival rate, relapse rate and treatment related

mortality using a set of classification techniques. The results show that the accuracy of our novel approach is approximately 3-4 % higher than that obtained using single feature selection models or without using any feature selection technique. The diagram of the entire process has been shown in Figure 1

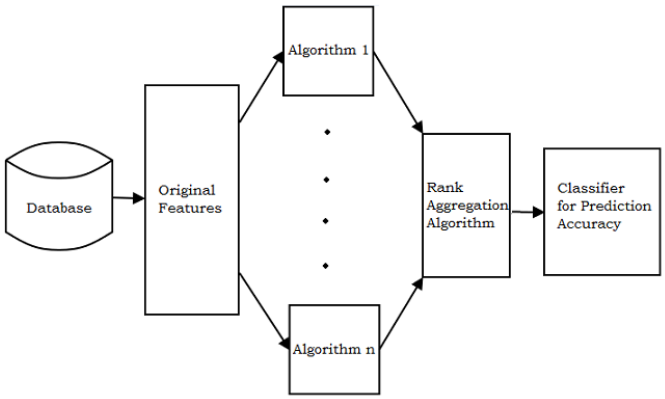

Fig. 1. Flow diagram of the entire process

B. Details

1) Feature Selection: Feature selection is a procedure used mainly in datamining for dimensionality reduction. Feature selection also has many benefits associated with it such as improving the prediction performance, detecting faster and cost effective predictors and in providing a better understanding of the process that generated the data [21]. In this paper we propose a novel ensemble feature selection technique using rank aggregation method which aims at giving a global ranking of features from the transplant data. In the first step we used five feature extraction algorithm using ranking and search method.The ranking method assigns ranks to each attribute based on the individual evaluation of each attribute. The entire model for feature selection is given in Fig 1. The feature selection base algorithms used are [22]

- Chi-Square—This algorithm evaluates features individually by measuring their chi-squared statistic with respect to the classes.
- Gain-Ration Attribute Evaluation—This algorithm evaluates features individually by measuring their gain ratio with respect to the classes.
- Info-gain Attribute Evaluation—This algorithm evaluates features individually by measuring their Information gain with respect to the classes.
- Symmetrical Uncertainty—Info-gain Attribute Evaluation—This algorithm evaluates features individually by measuring their symmetrical uncertainty with respect to the class.
- Filtered Attribute evaluation—an arbitrary attribute evaluator on data that has been passed through an arbitrary filter

By applying the above algorithms to features we generated five lists of ranked attributes on same data. Since we have five sets of ranks for each attribute, we then perform a rank aggregation mechanism to produce a global ranking of the attributes based on their significance level. We elected not to use feature selection techniques such as Principle Component Analysis or SVM because these algorithms scale the actual features in high dimensional space to produce synthetic features. Loosing the original features makes it difficult to interpreted the results in the medical domain where decision making relies on selection of original features.

2) Rank Aggregation: We propose an algorithm which uses a novel rank aggregation technique for assigning a global rank on the features which is uninfluenced by ranking algorithm biases. Rank aggregation can be done in various ways [23], [24]. The rank aggregation approach used in our algorithm is a modified version of rank aggregation method used for web searches[23], [24]. The mathematical formulation is shown in equation 1

$$Rank_{global} = \frac{1}{n}\sum_{i=1}^{n} Rank_n \tag{1}$$

where n is the number of list generated from n number of feature selection algorithms. $Rank_{global}$ is the global rank obtained after rank aggregation on the ranks obtained from n algorithms.

The benefit of using this approach is that, no prior knowledge about the contribution of each features or variables are needed since the actual rank produced by each feature selection algorithm is used as the contributing factor for our rank aggregation based feature selection algorithm.

Algorithm 1 Aggregated Ranking based Feature Selection

Input: Original Feature Set S
Output: Ranked Feature Set S'
STEPS :
1) Rank S using selected feature selection strategies
2) Sort the new ranked lists based on the index of the features in each list
3) Produce a matrix n x p of feature ranks where n is the index of features and p is the number of original feature selection algorithms used
4) Calculate the sum across each row for each feature in a separate column
5) Calculate the Global rank as given in equation (1)
6) Sort again based on rank in ascending order.This average is now our new global rank for each feature.
7) Select Top N features where N can be heuristically defined
8) The final list S' is the global ranked features.

The algorithm for rank aggregation based feature selection is given in Algorithm (1)

The output of this algorithm is a global rank for each feature. The significance of using a rank aggregation is that none of the feature ranks in the final list are biased due to specification of individual measures used for initial ranking.

Moreover, this global list represents a measurement of similarity between items in the various ranked lists apart from from actual rankings. The results of prediction accuracy shows comparable improvement in the favor of Rank aggregation over the individual ranking measures. Our approach uses merging of ranked lists where global rank is decided by the majority votes by ranking algorithms.

C. Data Set used

Data set used consisted of 1160 patients who received myeloablative, T-cellreplete, unrelated donor (URD) transplantation as treatment for AML. Transplants were facilitated by the National Marrow Donor Program (NMDP) between 1988 and 2006. DNA sample was obtained for each donor and recipient from the Research Sample Repository of the NMDP. Outcome data were obtained from the Center for International Blood and Marrow Transplant Research. Complete high-resolution HLA matching data at HLA-A, B, C, DRB1, and DQB1 were obtained from the NMDP retrospective typing program. A total of 121 attributes were studied. Gene expression data is binary (1- present and 0-absent). Response variables included treatment related mortality, leukemia free survival, relapse and death. The other variables were used to predict the outcomes above.

1) Preprocessing of the data: A preliminary domain based pruning was done on the data set to remove redundant (calculated) and missing values. The recipient KIR genetic variables were removed since previous analysis has demonstrated that they were not predictive of outcome after HSCT. [2]. The final data contained 1160 instances and 75 attributes including KIR genes, HLA allele matching at A, B, C, DRB1, and DQB1, age, race, sex, CMV status, graft type, Karnofsky score, disease status before transplant. Response variables used for prediction were

- Treatment Related Mortality—Indicator of death of patients which is only caused due to post treatment effects such as acute or chronic graft verses host disease which develops in a patient within a given period of time.
- Relapse Free Survival—Indicating whether the patient survived after BMT treatment without having a relapse, after a certain amount of time decided by the medical experts.
- Relapse Rate—indicating whether the patient had a relapse of AML

2) Evaluation: In order to evaluate the performance of our algorithm, we compared the classification accuracy between the prediction based on the feature subset produced using our rank aggregation technique with prediction based on the unprocessed features without using feature selection technique prior to model building for classification. The comparison of accuracy between the rank aggregation algorithm selected features and the results from the individual features have been considered in this paper as an additional evaluation criteria for the rank aggregated features. The mathematical formulation of accuracy measure A is given in 2

$$Accuracy = \frac{C}{C'} \tag{2}$$

where C is the Number of correctly classified samples and C' is the Total number of samples

3) Classification algorithms: The different classification algorithms which we used are—Decision Tree / AdTree, AdaBoost with Decision Stump or JRip, SMO, Logistic Regression, Voted Perceptron and Bayesian Network. The main reason behind using a variety of classification algorithms is to demonstrate the robustness of our approach. Our algorithm can be used along with a variety of Prediction measures including rule based, Bayesian Network. classification tree, ensemble based and even statistical measure like regression. We show in the result section how we get consistent prediction accuracy across all classification techniques considered.

IV. Experimental Results

In order to confirm the reliability of our overall data mining approach to accurately predict characteristics of a donor which are associated with improved outcome after HSCT for AML, we performed a comparative analysis of the prediction results of our algorithm with traditional statistical approach [2], [3] given in Table I. The Table I demonstrates the statistically significant variables that are selected by multivariate analysis with 95% Confidence Interval for relapse free survival after transplant. These are—Disease Status , Non-Caucasian, HLA matched/mismatched, Karnofsky, Performance Status, Age(categorical), Specific KIR Genes, Centromeric and Telomeric KIR groups, transplant (per year) and KIR B Content status. The features selected by our data mining approach shown in the first column of the Table shows that, our algorithm has been able to correctly capture the significant variables. All the statistically significant variables have been detected as the top ranked features with in top 15 by our rank aggregation algorithm. Moreover our algorithm also detected other important variables including—conditioning regimen during transplant, characterization of the AML as primary or secondary, donor and recipient sex match and graft source during transplant (bone marrow vs. peripheral blood derived stem cells). Our results show that our rank aggregation-based feature selection data mining algorithm could detect not only the previously identified statistically significant features, but also other novel features which had not been detected by any other approach. These results will direct physicians to explore other dimensions of donor characteristics which may have been overlooked. This is one of the several advantages of data mining; to detect hidden patterns which are not otherwise visible through human judgement or prior knowledge based variable detection.

Next, we analysed and evaluated the output of our rank aggregation algorithm. We classified the top 35 features selected by our rank aggregation algorithm to predict the survival rate, treatment related mortality rate, and relapse rate for patients with AML. A heuristic approach was used to determine the number of top features to select from the 75 ranked features.

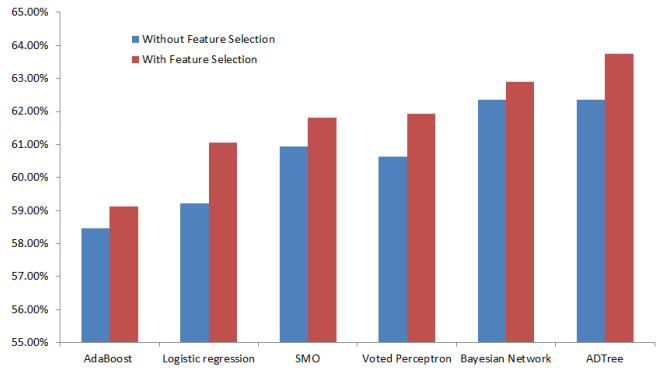

Fig. 2. Prediction Accuracy for treatment related mortality

Apart from observing the accuracy of our prediction , we also performed a comparative analysis of the result of classification on unprocessed data vs results obtained on using our rank aggregation algorithm. There is a striking 3-4% approximate overall improvement in the prediction. In figure 3 the prediction accuracy of survival rate is depicted. Our algorithm gives an additional 3-4% accuracy while predicting treatment related mortality. Similar trend can be seen in prediction accuracy comparisons for treatment related mortality and Relapse rate shown in Figure 3 and 4. These results shows that our algorithm gives a constant high accuracy for different kinds of classification algorithms as compared to when the features are used without applying our algorithm. Another, important factor of our rank aggregation algorithm is that, this approach is more robust as compared to when classifying with all the features. This algorithm is also scalable since the time complexity is in the order of $O(n)$ for the rank aggregation part of our algorithm. The classification algorithms has been used with Weka [22] which can handle huge amounts of datasets.

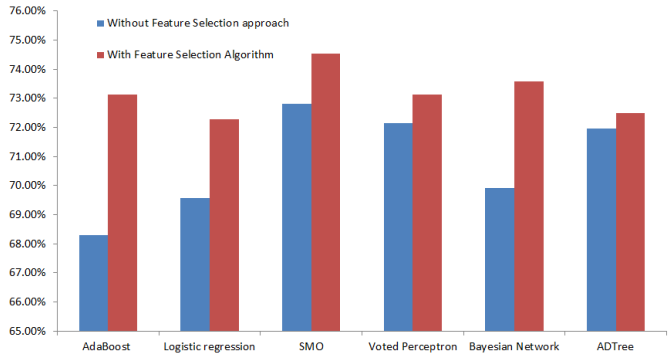

Fig. 3. Prediction accuracy of Survival Rate

In figures 2, 3 and 4 we can see that accuracy of our approach is almost consistent with different classification models. In contrast, the accuracy measure of classification based on the unprocessed features are fluctuating and model dependent. The main reason behind this, is the fact that our algorithm

TABLE I
CONFIRMATION OF FEATURE SELECTION RESULTS WITH MULTIVARIATE ANALYSIS OF THE DATA

Top 15 ranked Features (Rank aggregation based feature selection approach)	Statistical significance (95% Confidence Interval) based on previous studies	Description of the features
disstat	Significant	Status of disease at transplant Early, Intermediate, Advanced, Others
dnrrace	Significant	Race of donor
numhlaof10	Significant	Number of matches out of 10—based on HLA-A, -B, -C, -DRB1 and -DQB1
karnofpr	Significant	Karnofsky performance score for assessing patient's fitness
regi	Not Significant	Conditioning regimen during transplant
leuk2	Not Significant	Indicatior of wheher it is a primary or secondary AML case
dagecat	Significant	Age category of donor
Donor_ Neutral_ Better_ Best	Significant	Number of centromeric and telomeric gene content motifs containing B haplotype KIR genes
numtx	Significant	Total number of transplants the recipient has had
sexmatch	Not Significant	Donor and recipient sex match
graftype	Not Significant	Graft source : Bone marrow, PBSC, Cord blood
Donor Final Centro Grp	Significant	Centromeric KIR group—Cen A/A, Cen A/B, Cen B/B
Donor 2DS4 Length Groups	Significant	Presence or absence of Specific KIR Genes
Donor Final Telo Grp	Significant	Presence or absence of Telomeric group—Telo A/A, Telo A/B, Telo B/B

assigns a global rank which is not influenced and biased by individual feature selection models. Hence, robustness is preserved across all the classification models used. In Figure 5 we analyse the prediction accuracy from our rank aggregation based feature selection with that of individual feature selection models such as Chi Square based, Gain ratio , infogain , filter based and Symmetrical Uncertainty based feature selection models. This result shows that our approach gives a constant higher accuracy across a variety of classification algorithms than that of other single feature selection models. Overall, our approach gives a better result proving it has great potential in medical domain with significant benefits.

V. CONCLUSION

Data mining in the medical domain has been a successful approach for finding hidden patterns in the vast amount of patient-related data. Automated knowledge discovery can aid humans in medical decision making and in the identification of novel relationships and interactions between variables. In this paper we present a state of the art data mining approach to support donor selection for HSCT. We demonstrated that our rank aggregation algorithm can be used to efficiently select variables or features important to identify the optimal donor for HSCT. This entire approach has the ability not only to

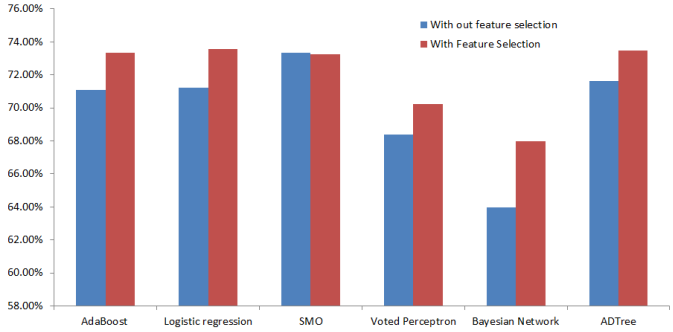

Fig. 4. Prediction Accuracy of Relapse Rate

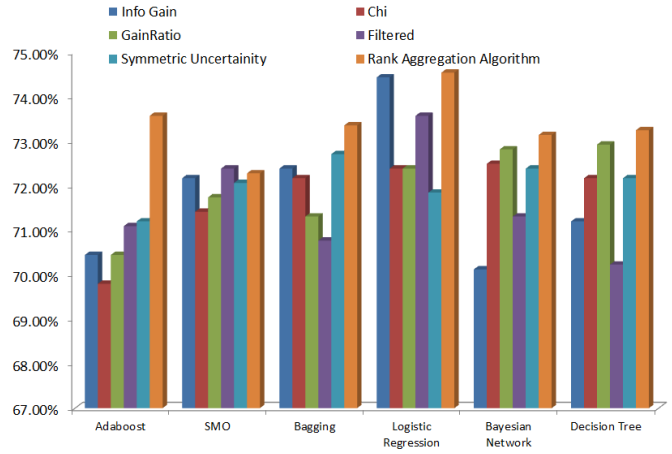

Fig. 5. Prediction Accuracy of Survival Rate for comparing rank aggregation algorithm vs single feature selection algorithms

indicate the significant features or characteristics of a donor, but also to eliminate those variables or features which are not reliable to predict the outcomes of interest. Moreover, our algorithm is robust on this large dataset and across a variety of the classifiers. Our future research will incorporate rule mining based feature evaluation techniques.

ACKNOWLEDGMENT

This work was supported by National Institutes of Health/NCI grant P01 111412, PI Jeffrey S. Miller, M.D, utilizing the Masonic Cancer Center, University of Minnesota Oncology Medical Informatics shared resource.

REFERENCES

[1] R. Storb and E. Thomas, "Allogeneic bone-marrow transplantation," *Immunological reviews*, vol. 71, no. 1, pp. 77–102, 1983.

[2] S. Cooley, E. Trachtenberg, T. Bergemann, K. Saeteurn, J. Klein, C. Le, S. Marsh, L. Guethlein, P. Parham, J. Miller, *et al.*, "Donors with group b kir haplotypes improve relapse-free survival after unrelated hematopoietic cell transplantation for acute myelogenous leukemia," *Blood*, vol. 113, no. 3, pp. 726–732, 2009.

[3] S. Cooley, D. Weisdorf, L. Guethlein, J. Klein, T. Wang, C. Le, S. Marsh, D. Geraghty, S. Spellman, M. Haagenson, *et al.*, "Donor selection for natural killer cell receptor genes leads to superior survival after unrelated transplantation for acute myelogenous leukemia," *Blood*, vol. 116, no. 14, pp. 2411–2419, 2010.

[4] M. Caligiuri, "Human natural killer cells," *Blood*, vol. 112, no. 3, pp. 461–469, 2008.

[5] C. Biron, K. Byron, and J. Sullivan, "Severe herpesvirus infections in an adolescent without natural killer cells," *New England Journal of Medicine*, vol. 320, no. 26, pp. 1731–1735, 1989.

[6] H. Shilling, N. Young, L. Guethlein, N. Cheng, C. Gardiner, D. Tyan, and P. Parham, "Genetic control of human nk cell repertoire," *The Journal of Immunology*, vol. 169, no. 1, pp. 239–247, 2002.

[7] G. Tourassi, E. Frederick, M. Markey, and C. Floyd Jr, "Application of the mutual information criterion for feature selection in computer-aided diagnosis," *Medical Physics*, vol. 28, p. 2394, 2001.

[8] B. Sahiner, H. Chan, D. Wei, N. Petrick, M. Helvie, D. Adler, and M. Goodsitt, "Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue," *Medical Physics*, vol. 23, p. 1671, 1996.

[9] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 8, pp. 1226–1238, 2005.

[10] I. Levner, "Feature selection and nearest centroid classification for protein mass spectrometry," *BMC bioinformatics*, vol. 6, no. 1, p. 68, 2005.

[11] T. Dietterich, "Ensemble methods in machine learning," *Multiple classifier systems*, pp. 1–15, 2000.

[12] Y. Saeys, T. Abeel, and Y. Van de Peer, "Robust feature selection using ensemble feature selection techniques," *Machine Learning and Knowledge Discovery in Databases*, pp. 313–325, 2008.

[13] M. Colonna and J. Samaridis, "Cloning of immunoglobulin-superfamily members associated with hla-c and hla-b recognition by human natural killer cells," *Science*, vol. 268, no. 5209, pp. 405–408, 1995.

[14] M. Uhrberg, N. Valiante, B. Shum, H. Shilling, K. Lienert-Weidenbach, B. Corliss, D. Tyan, L. Lanier, and P. Parham, "Human diversity in killer cell inhibitory receptor genes," *Immunity*, vol. 7, no. 6, pp. 753–763, 1997.

[15] N. Wagtmann, R. Biassoni, C. Cantoni, S. Verdiani, M. Malnati, M. Vitale, C. Bottino, L. Moretta, A. Moretta, and E. Long, "Molecular clones of the p58 nk cell receptor reveal immunoglobulin-related molecules with diversity in both the extra-and intracellular domains," *Immunity*, vol. 2, no. 5, pp. 439–449, 1995.

[16] N. Valiante, M. Uhrberg, H. Shilling, K. Lienert-Weidenbach, K. Arnett, A. D'Andrea, J. Phillips, L. Lanier, and P. Parham, "Functionally and structurally distinct nk cell receptor repertoires in the peripheral blood of two human donors," *Immunity*, vol. 7, no. 6, pp. 739–751, 1997.

[17] L. Ruggeri, M. Capanni, E. Urbani, K. Perruccio, W. Shlomchik, A. Tosti, S. Posati, D. Rogaia, F. Frassoni, F. Aversa, *et al.*, "Effectiveness of donor natural killer cell alloreactivity in mismatched hematopoietic transplants," *Science's STKE*, vol. 295, no. 5562, p. 2097, 2002.

[18] S. Giebel, F. Locatelli, T. Lamparelli, A. Velardi, S. Davies, G. Frumento, R. Maccario, F. Bonetti, J. Wojnar, M. Martinetti, *et al.*, "Survival advantage with kir ligand incompatibility in hematopoietic stem cell transplantation from unrelated donors," *Blood*, vol. 102, no. 3, pp. 814–819, 2003.

[19] S. Davies, L. Ruggieri, T. DeFor, J. Wagner, D. Weisdorf, J. Miller, A. Velardi, and B. Blazar, "Evaluation of kir ligand incompatibility in mismatched unrelated donor hematopoietic transplants," *Blood*, vol. 100, no. 10, pp. 3825–3827, 2002.

[20] R. Fagin, R. Kumar, and D. Sivakumar, "Efficient similarity search and classification via rank aggregation," in *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pp. 301–312, ACM, 2003.

[21] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[23] J. C. de Borda, *Memoire sur les Elections au Scrutin*. Paris: Histoire de l'Academie Royale des Sciences, 1781.

[24] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, "Rank aggregation methods for the web," in *Proceedings of the 10th international conference on World Wide Web*, pp. 613–622, ACM, 2001.