A Survey on Genetic Diseases Prediction

Pushkaraj Deshmukh, Shamika Thakur, Pooja Suryavanshi, Sufiya Shaikh, Viresh Vanarote

***Abstract*: Excesses of sex chromosomes results abnormalities in patients with the psychiatric diseases have lately been observed. It remains indistinct whether sex genetic material abnormalities are related to sex differences in some psychiatric diseases. An important group of genetic linkage with sex chromosomes is that the genes on X or Y chromosomes not only decide male and female traits but also carry many sex-related characters. While men are the only ones who inherit Y chromosome and Y- linked traits, both men and women can get X-linked genes since both inherit X chromosomes. In this survey, we will underline some genes located on X or Y chromosomes and recap their linkage with these psychiatric disorders. We will survey each of the genes and their link to specific diseases to predict the diseases which can occur in future.**

***Keywords*: Disease genes prediction, SVM, Naive Bayes, Classification.**

Introduction:

DNA abnormalities are the most common chromosomal abnormalities in humans. Sex chromosome aneuploidies can effect neuro development and often result in extra difficulties in psychological flexibility, constant attention, working, recall, verbal skill, and executive meaning impairment, while some of these symptoms partly cover with the phobia. Studies have well-known that genes in the sex chromosome may influence psychiatric disease by altering the basic partition process of the neurons, encoding proteins, and synaptic transmission and so on. In this survey, we gathered discovery in psychiatric study and discussed the relationship between sex chromosomes and phobia in arrange to give some helpful insights on sex specific inherited mechanisms for genetic disease prediction. A chromosomal disorder occurs when there is a change in the number or structure of the chromosomes.

This change in the amount or agreement of the genetic information in the cell may result in troubles in the growth, development and/or functioning of the body system. The incidence of chromosomal abnormality is approximately 1 out of 200 of a newborn. About 50% of first-trimester abortions 23 chromosomes from the mother and 23 chromosomes from the father. The percentages to be considered when it comes to human diseases, represents 1% of live birth, 2% of pregnancies in women older than 35, 50% of all spontaneous first- trimester abortions, responsible for more than 100 human syndromes and are more common than single-gene disorder. This can be a step by step pathway to incorporate more analytics and predictive algorithms for prediction of diseases.

Disease Prediction:

Earlier genetic diseases were predicted using genetic data of different generations, genetic data of many of their family members were required. It was time consuming. Carrier testing was also needed; it was only possible to predict if their family history was known. Large numbers of participants were required. Some disadvantages of genetic testing includes: Testing may increase anxiety and stress for some individuals. Testing does not eliminate a person's risk for disease. Results in some cases may return inconclusive or uncertain results. Waiting for test results can be very nerve racking and more important the correct results depends on the predictive technologies used. If the diseases are predicted correctly then only the proper treatment can be induced further. A major drawback of a genetic risk prediction test for common diseases is that it can't be used as a diagnostic instrument because it has no accuracy. Over the past 10 years. Accuracy of genetic risk prediction algorithms for common diseases has improved and need to be improved, but due to complex nature of common diseases the genetic prediction algorithms will never be entirely accurate.

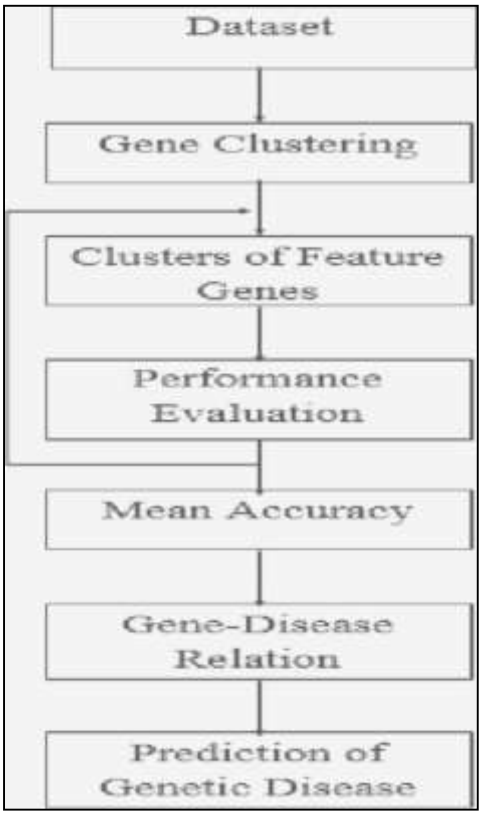

Fig.no.1: Flow of Prediction of Genetic Disease

Digital advancement in healthcare industry is expected to be achieved with the help of different domains as per the requirements. The prediction of diseases is a concern due to lack of resources in the medical field, which can be solved using suitable technology support in this regard and it can be highly beneficial to the medical fraternity and patients.

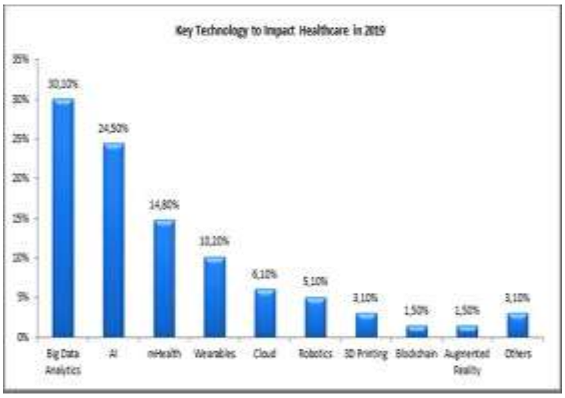

Fig.no.2: Key Technology to impact healthcare in 2019

The above graph depicts the survey on the technologies that can challenge the traditional, reactive approach to healthcare.

System Architecture:

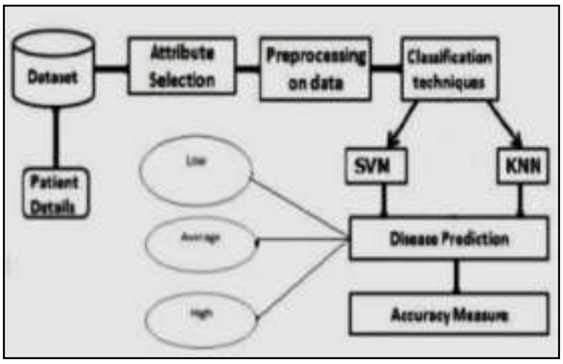

Fig.no.3: Accuracy Measures

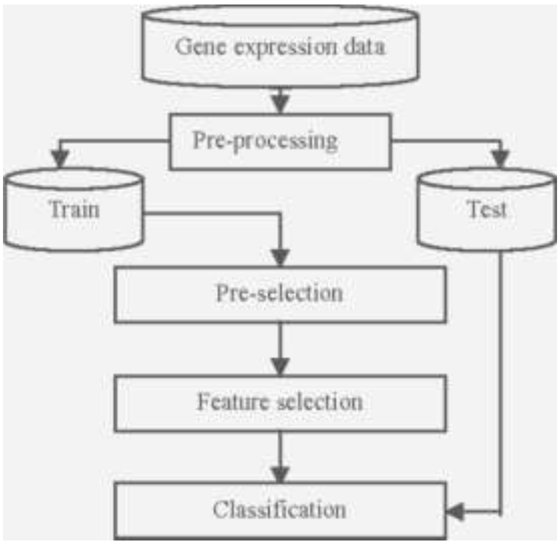

Fig.no.4: Classification

Genetic Disease Prediction:

The early disease prediction is done using algorithms like KNN, NB. Later feature selection algorithms like LASSO have been used to select useful attributes for prediction, After selection process, they are further selected, cleaned and made into desired form. The classifier performance was checked on pre –processed data (selected features) using K-fold cross validation method. The disease prediction will be trained on the dataset of diseases to do prediction accurately and produce confusion matrix, which was used in performance evaluation matrices to check the performance of the classifiers. Data consist of structured data as well unstructured data structured data: Patient's basic information e.g. age, gender, life habits etc. Unstructured data: Patient's narration of his/her illness, the doctor's interrogation records and diagnosis (mostly text data).

Algorithm:

1. **Support Vector Machine:**

Input: Set of (input, output) training pair data; call the input feature as x and the output result as y. normally, here can be a lot of input features xi (i.e. x= [XXX, XYY, XXY, YYY, XYY, YXX, YY X], [45 pairs, X,
Y], [XXX, XXY] etc.). Output: In this algorithm, first we generate ARFF File classifier and this classifier file loads all instance data and attribute value i.e. user input. If system accepts the valid input, SVM classifier calls the file data and starts the classification as per requirement. The ARFF file attributes decide which disease accuracy is high and predicts the final output. Output as Y = Trisomy, Turner Syndrome, Klinefelter Syndrome, Pentasomy X disease.

2. **Naive Bayes:**

Input: Set of (input, output) training pair data; call the input sample features as x and the output result as y. typically, there can be lot of input features xi. (i.e. x = Intellectual Disability, Facial Appearance, Kidney problem etc.) Output: In this algorithm first, we generate Arff File classifier and this classifier file loads all instance data and attribute value i.e. user input. If system accepts the valid input, SVM classifier calls the file data and starts the classification for each requirement. ARFF file attributes decide which disease correctness is high and predicts the last output. Output as Y = Trisomy (XXX, XYY, XXY, YYY, XYY, YXX, YYX), Turner
Syndrome(X, Y), Klinefelter Syndrome (XXX, XXY), Pentasomy X disease (XXXXX). Digital advancement in healthcare industry is expected to be achieved with the help of different domains as per the requirements. The above graph depicts the survey on the technologies that can challenge the traditional, reactive approach to healthcare.

Accuracy:

Commercial labs often give faster results (usually within 2-4 weeks) than research centers (minimum of 4 weeks or often longer). Up to 40% of at-home genetic test results may be “false positive”. Our project accuracy is average 87% to 96%, the sum average accuracy counts to 91% to 96%.

Algorithm Used	Accuracy
SVM	45.33%
Naïve Bayes	52.30%

Table 1: Accuracy for medical diagnosis - Heart Disease Prediction

Future Work:

Besides the current genetic prediction system using DNA, the traceability of pathogenic agents in the human body through molecular analysis is also a field to be further exploited. In the future it can be extended to perform genetic analysis for any genetically encoded feature of a person to diagnose current illness, predict future disease risk and to define other less medical relevant traits.

Conclusion:

It is possible to predict diseases using machine learning techniques. Earlier methods of predicting genetic diseases were time consuming and complex. There is a need of simpler methods for increasing accuracy.

References:

[1] K. G. Becker, K. C. Barnes, T. J. Bright, and S. A. Wang, “The genetic association database.” Nature Genetics, vol. 36, no. 5, pp. 431–432, 2004.

[2] D. Botstein and N. Risch, “Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease.” Nature Genetics, vol. 33, no. 33 Suppl, pp. 228–237, 2003.

[3] X.Wu, R. Jiang, M. Q. Zhang, and S. Li, “Network- based global inference of human disease genes,” Molecular systems biology, vol. 4, no. 1, p. 189, 2008.

[4] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, “Associating genes and protein complexes with disease via network propagation.” Plos Computational Biology, vol. 6, no. 1, p. e1000641,2010.

[5] Y. Li and J. C. Patra, “Genome-wide inferring gene– Phenotype relationship by walking on the heterogeneous network,” Bioinformatics, vol. 26, no. 9, pp. 1219–1224, 2010.

[6] U. M. Singh-Blom, N. Natarajan, A. Tewari, J. O. Woods, I. S. Dhillon, and E. M. Marcotte, “Prediction and validation of genedisease associations using methods inspired by social network analyses,” PloS one, vol. 8, no. 9, 2013.

[7] R. M. Piro and C. F. Di, “Computational approaches to disease gene prediction: rationale, classification and successes,” Febs Journal, vol. 279, no. 5, pp. 678–696, 2012.

[8] M. A. V. Driel, J. Bruggeman, G. Vriend, G. B. Han, and J. A. M. Leunissen, “A text-mining analysis of the human phenome,” European Journal of Human Genetics, vol. 14, no. 5, pp. 535–542, 2006.

[9] S. K¨ohler, S. Bauer, D. Horn, and P. N. Robinson, “Walking the interactome for prioritization of candidate disease genes,” American Journal of Human Genetics, vol. 82, no. 4, pp. 949–958, 2008.

[10] N. Natarajan and I. S. Dhillon, “Inductive matrix completion for predicting gene–disease associations,” Bioinformatics, vol. 30, no. 12, pp. i60–i68, 2014.

[11] S. Dieleman and B. Schrauwen, “Deep content- based music recommendation,” in International Conference on Neural Information Processing Systems, 2013, pp. 2643– 2651.

.