3.1 Research Design

Over the last few decades research in biological and medical science has witnessed an increase in the usage of artificial intelligence based techniques for designing various tools and techniques that help to achieve better patient care. This includes techniques that help in better prognosis / diagnosis of diseases, providing expert systems for consultation where human medical experts are not available, providing intelligent screening systems for early and timely detection of diseases, analysis of survival patterns of diseases etc. The use of Artificial intelligence in medicine has shown substantial progress in reducing the workload on the clinicians by providing decision support tools, detecting artefact with the ability to learn and adapt on-line. The overall objective of this endeavour is to study the application of artificial intelligence in prognosis of cancer especially screening and diagnosis of cervical cancer.

The basic objectives of the study are:

I. To study the existing methods and algorithms in artificial intelligence / machine learning and understand their working, advantages, disadvantages and application. To analyze and identify which among these algorithms can be applied in the domain of medical diagnosis.

II. To perform a detailed study of diseases with high mortality rate with the help of a medical specialist of the concerned field, in terms of causes, symptoms, etiology, risk factors, statistics, screening practices, pathological tests etc.

III. To collect, preprocess, analyze, organize and develop a primary database for training, testing, validation and implementation of artificial intelligence based algorithms in solving different problems for prognosis / screening of diseases.

IV. To analyze, understand and compare these algorithms on different parameters such as performance, true positive rate, false positive rate, true negative rate, false negative rate, receiver operator curve characteristics, robustness, quantity of training data required, and validity across different datasets etc. As such analyze
and study the performance difference i.e. what makes an algorithm perform better than other counterparts.

V. To propose a prognostic framework that could potentially aid a medical specialist in performing expert diagnosis. To test the proposed framework and verify its authenticity with the help of specialist doctors in the concerned field.

3.1.1 Various Considerations

In order to achieve the above marked objectives in an efficient manner, some research considerations have been worked out. These considerations are:

I Investigation of all the diseases in medical science is a herculean task for a single research to accomplish, keeping in view the vast dimensions of the field of medicine and therefore, the efforts required. For this reason, it was considered to select a single medical disorder which has a high incidence and mortality rate and design, develop and test the artificial intelligence based prognostic framework for it in such a way as could be easily modified for implemented for many other disorders as well.

II For training, testing and validation the artificial intelligence based algorithms on the selected medical disorder; secondary sources of data were considered on the initial phases of study. As the study advanced, the need for more detailed and precise data was required. Consequently, it was decided to obtain & organize data from primary sources. This was followed by processing the same to be in the shape of feeding it to the various algorithms under study. The primary data was obtained from multiple healthcare centres which were in the shape of slides of pathological tests. For extracting the digital data out of the medical slides, they were subjected to the examination under high resolution microscope and followed by capturing images of objects of interests. The analysis of slides under high resolution microscope mounted with digital cameras was carried out at the laboratories of Department of Pathology, Government Medical College Jammu, Department of Zoology and Department of Environment Science in the parent university.
Extracting and organizing digital information from the medical images was a very cumbersome task which took a lot of time to complete. For profiling these medical images an open source utility written in Python called CellProfiler was used.

For consolidation of data, design of conceptual and physical schema, analysis and interpretation of information in database tools like MS Excel spreadsheet, and MS Access were used. For programming and development of intelligent prognostic tool, plotting graphs and drawing conclusions, MATLAB (Matrix laboratory) was used.

3.2 Research Methodology

3.2.1 Research Phases

The entire work was divided into eight phases so that the desired objectives could be efficiently achieved. A brief description of each phase is given below:

3.2.1.1 Phase-I

In this phase, a detailed study of various medical disorders which have a noted high incidence and high mortality rate was carried out. Also various artificial intelligence based algorithms were studied in terms of their advantages, disadvantages and areas of application. Out of the various diseases, diabetes type-II (hyperglycaemia) was selected for pilot study. After a detailed literature study about the selected disease followed by consultation with the medical expert, ten different physiological parameters were identified which play an important role in manipulation of the disease. These parameters were Age, Family history, Weight, Sex, Drinking Smoking, Thirst, Frequency of urination, Height and Fatigue. On the basis of these parameters a rich primary database was constructed which contained data about 400 people from across the length and breadth of the state. To ensure variety and diversity in the database care was taken to include people from different age-group, cultural, economical and geographical
background. Once the database was populated with enough number of tuples, it was used to train, test and validate various artificial intelligence and statistics based algorithms, which include K-Nearest, Neighbour, Naive Bayes and Artificial Neural Networks. Once the system was trained using the intelligent algorithms, the test data were used to evaluate their working efficiency in prediction. The reliability of the system was evaluated by computing the mean absolute error between the predicted values and actual values in the test data. After the application of the different algorithms to same data set, it was found that among all the algorithms under study, ANN performed the best prediction with an accuracy of about 96% followed by Naïve Bayes and KNN having an accuracy of about 95 and 91%, respectively. The results of the pilot study suggested that these algorithms can perform good prediction with least error in the field of medical diagnosis and can be used to build various tools for prognosis and thus can supplement medical doctors in performing expert diagnosis. One of the limitations of this study is that the database constructed for the medical disorder under consideration was limited to only physiological parameters, which can be improved to include clinical parameters also. This may further enhance the working efficiencies of the algorithms under study.

The pilot study of application of multiple machine learning algorithms for prognosis of type-II diabetes provided us with an insight into the relative ability of artificial intelligence based algorithms for solving problems in the field of medical diagnosis.

3.2.1.2 Phase-II

Motivated by the encouraging results obtained in the pilot study, a fresh disease which was rather more complex for diagnosis with higher incidence and mortality rate was selected for further investigation. Cervical cancer is the fourth most common cancer in women and the seventh overall, in the world. It being one of the leading causes of female mortality due to cancer in the world; was selected for developing some artificial intelligence based automated / semi-automated tool for its preliminary screening. A thorough study of the concerned medical literature was carried out with special reference to the pathological tests done for screening and diagnosis of cervical cancer. The Papanicolaou test (Pap smear) has been the widely used method in cervical cancer
screening for many decades and has showed a dramatic lowering of incidents of cervical cancer and hence in related mortality rates in many countries. In taking a Pap smear, cells are scraped from the outer opening of the cervix for microscopic examination and to lookup for irregularities. The aim of the test is to detect any pre-cancerous or potentially precancerous alterations called cervical intraepithelial neoplasia (CIN) or cervical dysplasia. Pap test is also used to detect endocervix and endometrium abnormalities and infections. In many developed countries, regular Pap smear screening is highly recommended for females who have had frequent sex with multiple partners. If any unusual findings are observed the test may need to be repeated within a year. If the abnormality observed requires closer examination, a detailed cervical inspection by colposcopy may be done. HPV DNA testing may also be suggested to such patients, which acts as a supplementary to Pap smear testing. Once the sample is obtained, Papanicolaou technique is used to stain it. Staining using this technique helps to differentiate the cells in smear preparation from various other bodily secretions as unstained cells cannot be seen under a simple compound microscope. Most of the abnormal results are mildly abnormal (called low-grade squamous intraepithelial lesion (LSIL)) which indicates HPV infection. Most low-grade cervical dysplasia relapse on their own without usually causing cervical cancer, but presence of dysplasia can act as a warning that greater monitoring is needed. Generally, some Pap results are high-grade squamous intraepithelial lesion (HSIL), and very few of them indicate cancer. After detained discussion with the medical experts of the concerned field, thirty nine (39) clinical parameters of the cells of cervix which play a deciding role in prognosis of cervical cancer were identified. Among the 39 selected features, 15 correspond to the measurement of the size and dimension whereas the other 24 correspond to measurement of shape and contour of the cytoplasm and nucleus of the cell. The features which were selected were:- Area of the Cytoplasm, Center of Cytoplasm X-axis, Center of Cytoplasm Y-axis, Compactness of Cytoplasm, Eccentricity of Cytoplasm, Extent of Cytoplasm, Form factor of Cytoplasm, Major Axis Length of Cytoplasm, Maximum Radius of Cytoplasm, Mean Radius of Cytoplasm, Median Radius of Cytoplasm, Minor Axis length of Cytoplasm, Cytoplasmic Orientation, Cytoplasmic perimeter, Solidity of Cytoplasm, Location of Center of Cytoplasm X-Axis, Location of Center of Cytoplasm X-Axis,
Maximum Farthest diameter of Cytoplasm, Minimum Farthest diameter of Cytoplasm, Area of the Nucleus, Center of Nucleus X-axis, Center of Nucleus Y-axis, Compactness of Nucleus, Eccentricity of Nucleus, Extent of Nucleus, Form factor of Nucleus, Major Axis Length of Nucleus, Maximum Radius of Nucleus, Mean Radius of Nucleus, Median Radius of Nucleus, Minor Axis length of Nucleus, Nuclear Orientation, Nuclear perimeter, Solidity of Nucleus, Location of Center of Nucleus X-Axis, Location of Center of Nucleus X-Axis, Maximum Farthest diameter of Nucleus, Minimum Farthest diameter of Nucleus and Nucleus-Cytoplasm Ratio.

3.2.1.3 Phase-III

In this phase data collection was made for developing a rich primary database of cervical cells. This database could serve as a potential tool for designing, developing, training, testing and validating various artificial intelligence based systems for prognosis of cervical cancer. Various sources from where the data could be obtained were identified which was followed by manual collection of the data from them. The data collected from these primary sources was converted into the digital form and as such given a shape as could be usable for machine learning analysis.

3.2.1.4 Phase-IV

In this phase various artificial intelligence based algorithms were applied on the primary database developed in pervious phase, followed by analyzing and comparing the working efficiencies of these algorithms. From the database prepared, ten different sets of training data, testing data and validation data were derived out for realization of ten-fold cross validation. For realization of various artificial intelligence based algorithms, programming codes were written for some of the algorithms in MATLAB (Matrix laboratory) while for others WEKA workbench (version 3.6.10) was employed. WEKA is a machine learning workbench, which is written in Java and consists of a collection of data analysis algorithms useful for designing predictive models as solutions to the Real-world problems. The comparisons of working efficiencies were made for ten-fold cross validation on the basis various performance metrics including specificity, sensitivity and Receiver-Operator curve analysis.
3.2.1.5 Phase-V

Among all the algorithms implemented in previous phase, the best performing algorithms were identified, for auxiliary analysis to further enhance their recantation rate or true positive rate. Further in this phase a Hybrid Ensemble technique is proposed, which works by constructing a large number of classifiers at training time and producing the class that is the mode of the classes output by individual classifiers. The method improves the predictive performance by coalescing all the classifying potential of individual classifiers. The final classification, as output is selected by taking the votes for particular instance from various individual prediction models taken under consideration. By using ensemble method the likelihood of misclassifying a particular instance are reduced considerably, and thus provides a greater confidence to the classification process. If an instance is wrongly classified by some classifier, the error is corrected by the right classification done by other algorithms. Thus auto correcting of errors is achieved by nullifying the chance of wrong final classification by a classifier. The ensemble system considered is developed using fifteen different classification algorithms. The algorithms considered were Random sub space, Radial basis function network, Multiclass classifier, Random forest, Bagging, Rotation Forest, Decision Tree, Ensemble of Nested dichotomies (END), Decorate, K-NN, Support vector machine, Filtered Classifier, Decision Table, Back propagation artificial neural network, and Naïve Bayes. These algorithms were considered owing to their noted good performance for classification of complex datasets. The system can be viewed as an ensemble of ensemble classifiers as internally bagging, random forest, rotation forest, random committee, random subspace, decorate and nested dichotomies themselves are based on ensemble principle.

3.2.1.6 Phase-VI

In this phase, a novel frame work was designed and developed for screening of cervical cancer. The framework was based on the methodology of Divide and Conquer technique for simplifying the problem at hand. It worked by the implementing the backpropagation neural network in three sub phases or modules. The first sub-phase is concerned with application of backpropagation neural network for segregation of the test
case as normal or abnormal smear. The other two phases are concerned with detailed classification of normal and abnormal cases into respective classes as per the latest 2001-Bethesda system of classification of Papanicolaou smears. As such the test case is finally classified into one of the seven classes of the Papanicolaou smears image analysis.

3.2.1.7 Phase-VII

This phase is concerned with research findings, conclusion, recommendations and scope for further research.