Paper to be published in Artificial Intelligence in Medicine --- Elsevier…

**Title:** 10 to 13 words..

**Summary / Abstract 250/300 words:** Objective / Methods and materials / Results/ Conclusion…

**Introduction:** Apnea / Data prepossessing and feature selection with / Artificial Neural network / Evaluation technique / Literature review with gaps / What we are doing (problem definition) / What we are doing objectives and methods / How paper is organized..

(Check good introduction from other papers and see how it is written)….

**Previous Research:** literature review … Shekar to include the part here what I have told you..

As Machine Learning has matured, new opportunities have emerged for various applications in the Medical domain. Neural networks have been established at the forefront of applications in these domains, for medical disease diagnosis.

A decision-reference system was proposed by Vanisree K et al for diagnosis of Congenital Heart Diseases[5] where a Multi layer feedforward network (a type of neural network) was trained on a benchmarked dataset. Various physiological features of a patient were considered in this work, including signs, symptoms and medical test parameters. This system managed to achieve an accuracy of 90% in providing an accurate measure whether the patient was diseased or not.

In another work by C.S Dangare et al [7], a Heart Disease Prediction System was developed based on neural networks. The system predicts the probability of a patient developing a heart disease based on 13 continuous medical parameters like blood pressure, cholesterol and discrete features such as obesity and health habits such as smoking. The study manages to diagnose heart disease with 99.25% accuracy by training a relatively simple neural network with 1 hidden layer on 570 training examples. This study is a good example of how effectively neural networks even in their nascent form, can assist in diagnosis of medical diseases.

An automated speech recognition (ASR) system was developed by Krzysztof Wołk et. al [8], inspired by EU-BRIDGE project, intended to automate transcription and translation technology, and cross the language-barrier between medical communities worldwide, promoting sharing of medical research. Further, access to translations of a foreign patient’s medical data could save their lives.

This work utilized a sophisticated recurrent neural network with 750 hidden layers trained on sentences of Polish-to-English and vice-versa, composed of 50 words or fewer with 500,000 iterations and manages to achieve a working accuracy under strict hardware constraints.

This is fine try to find out the following points from the literature: You can even make a table with following:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Authors name | Methodology adopted( Different types of topology and variations on neural network) | Applications used……. | Advantages / Disadvantages of the approaches if any | Data Exploration techniques or any preprocessing technique used as well as any major modification used.. |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

Last you can write the all the gaps you have identified in the literature.

1) Neural network methodology with objectives

2) Different applications where it is used and how

3) Advantages and disadvantages of particular approach

4) Data Preprocessing techniques used.

5) Gaps in all the papers

**Method:**

Data source: In order to perform the research reported in the manuscript, 367 observations of new born babies were collected from NICU, Kasturba Hospital. The mean Gestation age of all the babies was 29.56 diagnosed with…..Ethical approval has been obtained from Institutional Ethics committee of Manipal University. The anonym zed data set obtained consists of more than 100 predictor variables with outcome variables pertaining to number of apnea episodes from the time of admission to discharge. The variable which signifies the number of apnea episodes from Day 4 to Day 7 was chosen to be the final outcome variable and the rest were discarded. We chose this because we want to predict whether apnea persists till the 7th Day based on data collected during the first 3 days. This was made into a binary classification problem by transforming the outcome column where the number of episodes greater than 0 would indicate Presence of apnea and the number of episodes equal to 0 would indicate Absence of apnea. With the help of medical experts advice only 20 variables taken during the first 3 Days are been used to predict the presence or absence of apnea from Day 4 to Day 7. Table 1 and Table 2 show the list of predictor variables and the outcome variable of the final data set respectively. ( I feel we should include this part in data nalytics otherwise it doesn’t have any sence)

Table 1. List of Predictor Variables ***[ TO ADD REAL NAMES]***

|  |  |
| --- | --- |
| **Predictor Variables** | **Description** |
| Gestation Age | Numeric |
| AGASGA | Categorical |
| Surfactant | Dichotomous Categorical |
| Birth Weight | Numeric |
| Steroids | Categorical |
| Head circumference at birth | Numeric |
| Birth cry | Dichotomous Categorical |
| Apgar Score at 1st minute | Numeric |
| Apgar Score at 5th minute | Numeric |
| Need of Resuscitation | Categorical |
| Mode of Resuscitation | Categorical |
| Desaturation1\_3DAYS | Numeric |
| Brady1\_3DAYS | Numeric |
| BD1\_3DAYS | Numeric |
| BD4\_7DAYS | Numeric |
| IMVDAYS | Numeric |
| NIVDAYS | Numeric |
| ECHO | Categorical |
| PDA | Dichotomous Categorical |
| DAYS OF SUPPORT | Numeric |

Table 2. Outcome Variable

|  |  |
| --- | --- |
| **Outcome variable** | **Description** |
| Apnea | Dichotomous Categorical |

Data understanding and preparation: Data analysis and preprocessing concept include here…Genetic / PCA and other preprocessing include here…SMOT also include here method and why we need to use it. Mention about methods and how we have used them in detail….

The data set to be preprocessed consist of 21 variables and 367 observations. All the categorical variables were converted to numeric codes where each code represented a particular category. The missing values of numeric columns were handled by replacing the missing value with the mean of all values corresponding to the similar group particular number of apnea episodes from Day 4 to Day 7.(this should be in general and provide reference). Similarly, the missing values of categorical columns were handled by replacing the missing value with the mode of all values corresponding to the particular number of apnea episodes. For some categorical columns such as Mode of Resuscitation and Steroids, where there were many missing values, a new category or class representing missing values was created rather than replacing the missing value with the mode or dropping the rows [ ]. This was done because the reason so many values were missing could be semantic rather than bad record keeping and they could represent some information crucial for predictive analysis. (put it in better sentence no meaning). The observations consisting of higher number of missing values were removed. The final data set comprised of 364 observations. Representation of categorical variables as integers is not logically correct (WHY? Reference []). For example, the column Steroids is represented by 4 numeric codes shown in Table 3.

There were many missing values in the data set but after summarizing and analyzing the entire data set with the help of scatterplots and histograms, no outliers were found. (Doesn’t have any meaning please tell how?)

Table 3. Representation of “Steroids”

|  |  |
| --- | --- |
| **Numeric Code** | **Description** |
| 1 | Not Given |
| 2 | Partially Given |
| 3 | Completely Given |
| 4 | Not known |

In Table 3, the numeric codes for the categories are not logically correct as for machine learning models “Completely Given” having the code 3 would be 3 times that of “Not Given” having the code 1 and hence be logically greater in value than it. To have a logically correct representation of categorical variables, all of them were dummy coded with 1’s and 0’s as seen in Table 4.

Table 4. Dummy coding of “Steroids”

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Numeric Code** | **Description** | **Dummy Code** | | |
| **Not Given** | **Partially Given** | **Completely Given** |
| 1 | Not Given | 1 | 0 | 0 |
| 2 | Partially Given | 0 | 1 | 0 |
| 3 | Completely Given | 0 | 0 | 1 |
| 4 | Not known | 0 | 0 | 0 |

In Table 4, the numeric codes for Steroids have been converted to Dummy codes which is a more logical representation. For any column with k categories, the codes can be represented with k-1 categories. In table 4, “Not known” can be represented as 000 rather than 0001 with an extra column which is more redundant. Having k-1 categories not only reduces redundancy but also has a lesser dimensionality. Similar dummy coding was done for all categorical variables and the total number of columns increased from 21 to 32. The final data set to be used for analysis had 364 observations and 32 columns. Out of the 32 columns, 31 were predictor variables and 1 was a binary outcome variable. Since the entire data set consists of only numbers, it could be successfully normalized to 0-1 scale. Normalization of this data set was really crucial because it had to be brought to the same scale for optimization when applied to the machine learning algorithms []. (Simple sentence why normalization is done in general0. Min-Max Normalization was used to bring the entire data set to the 0-1 scale.

To assess (or to check) the medical importance of the variables chosen for prediction of apnea, and to shortlist the variables out of the entire list, analysis of data was done for both numeric and categorical variables (Confusion is there first we are doing it through medical expert in first para the are we going to check them with statistically?). Some (which one) visualizations were used to spot trends in the data set. No outliers and redundant data were found in the list and the analysis was done after filling in the missing values as done in the preprocessing step. To analyze the medical importance of each of the variables used, bar charts were used to represent categorical variables and scatterplots were used to represent numeric variables.

The relationship between each of the numeric variables was checked with the outcome variable. The scatterplots with smoothing curves shown from Fig.(a) to Fig. (l) represent this relationship of the variables chosen in the final data set. The smoothing curves of the following figure show that the variables used for analysis are distinctive in nature. They are plotted against the probability of the occurrence of apnea. The numeric predictors are both indicative of presence as well as absence of apnea. Variables like Apgar score at 1st Minute and 5th Minute are more densely populated in the region that has low probability of presence of apnea or a high probability of absence of apnea. Variables like Desaturation from Day 1 to Day 3, Days Of Support and NIV Days seem to be good predictors of presence of apnea. When the Desaturation value is between 5 and 15, presence of apnea has very high probability, making it a good predictor of apnea. Similarly, when the Days Of Support value is between 40 and 60, there is a good chance of presence of apnea making it a good predictor too.

|  |  |  |
| --- | --- | --- |
| (a) | (b) | (c) |
| (d) | (e) | (f) |
| (g) | (h) | (i) |
| (j) | (k) | (l) |

The analysis of numeric data as shown previously was done using scatterplots and variables like Days of Support, Desaturation from Day 1 to Day 3, NIV Days were considered as good predictors of apnea. The analysis of categorical variables in the data set was done using filled bar charts where each category was analyzed to visualize the ratio of presence and absence of apnea as shown from Fig.(a) to Fig(h). From these visualizations, these categorical variables seem to be good differentiators of presence and absence of apnea. Need of Resuscitation, Mode of Resuscitation, Steroids and Echo seem to be good predictors of presence of apnea.

(How we are using this concept ie we got some thing but now we should relate this is actual research work)

(We can make this together with medical expert to select variables from 100 what we have said na before) or else we can say to verify the selected variables usefull ness in research)

(Put this diagrams in a proper format with proper tag line explaining what it is)

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

(Put this diagrams in a proper format with proper tag line explaining what it is)

Prediction models: Neural network and all its variant explain here which we are using…..Ensemble approach / Holdout method / K fold Cross validation / Evaluating model performance measure /

**Results: Put the result in tables with all evaluating performance measure we have defined above…Tables graphs…atleast one or two ROC curves..**

**If possible put all ROC curves in one graph with AUC measure**

**Discussion:** Discuss the results and algorithm, slowly come to conclusion why and which algorithm is better and its drawbacks if any Check out in papers what discussion contains and how it is different from results and conclusion….

**Conclusion with future work:** We have to conclude the paper with something novel interesting and with future work…