Flu Shot Learning: Predicting H1N1 and Seasonal Flu Vaccination

Jorge Pais, José Baptista and Pedro Duartee
up201904841@edu.fe.up.pt; up201904814@edu.fe.up.pt; up201905050@edu.fe.up.pt

***Abstract*—**

I. Introduction

Pandemics have rarely taken center stage in the way they have recently with COVID-19 in 2020. Vaccines are a key public health measure used to fight infectious diseases like COVID-19. They provide immunization for individuals, and enough immunization in a community can further reduce the spread of diseases through ”herd immunity.” In 2009, the H1N1 influenza virus, also known as swine flu, caused a global pandemic that was estimated to have resulted in 150,000 to 600,000 deaths worldwide. A vaccine for H1N1 became available in October of that year. In this study, a machine learning model was developed to help estimate the probability of a person receiving seasonal and H1N1 vaccines. For this, several classification methods were explored and compared between each other in order to figure out which one exhibited the best performance.

II. Data Resources

For the competition partaken in this project, the dataset was provided by DrivenData, and it comes from the National 2009 H1N1 Flu Survey (NHFS) which was collected through telephone interviews. The dataset consists of 36 attributes, varying from numerical with both ordinal and binary variables, and also categorical attributes. For the training data, there were 2 labels associated with each respondent, indicating whether or not each respondent had taken the H1N1 and Seasonal flu vaccines.

A. Data Cleaning and Pre-processing

The first step in the data analysis was to check if the data has any duplicate (i.e. duplicate respondent ids) and missing values, duplicates were not observed, but the dataset had many missing values in different attributes so, one of the first concerns was to clean the data. Some the attributes, namely `employment concern` and `employment occupation` had the highest missing data (13470 values). The *Pandas* library in Python identifies these missing values as `NaN`, and this can be identified and handled by different imputation classes from *SKLearn*. Another problem is that some of the data is categorical, which can be solved simply by encoding each possible category within a feature.

B. Feature Correlation

The next step involved checking of correlation between the attributes. This was done using the *Seaborn* library's correlation heatmap. It was noted that some attributes exhibited mild positive correlation, with two stand-out attributes named `doctor recc H1N1 vaccine` and `doctor recc seasonal vaccine`, which were positively correlated by 60%.

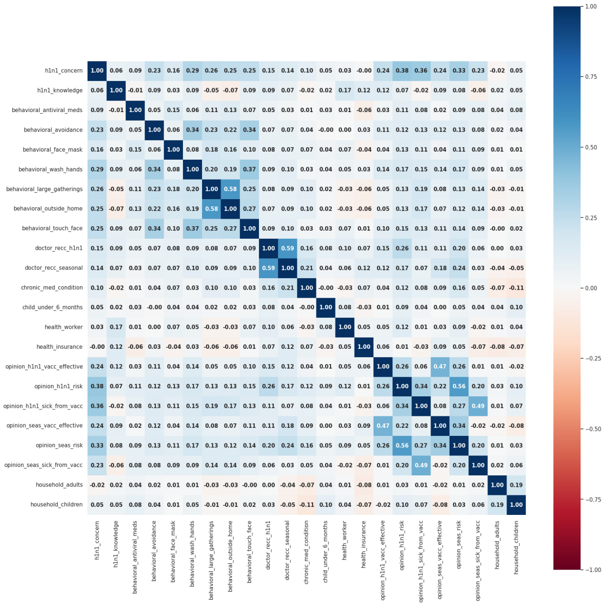

Figure 2.1 - Feature Correlation Heatmap

C. Class Balance and Label Correlation

Observing the distribution of the two target variables, shown in Figure 2.2, approximately half of individuals have been vaccinated for the seasonal flu, while only 20% have been vaccinated for H1N1. As for class balance, we can say that the distribution of individuals vaccinated for the seasonal flu is balanced, while the distribution of those vaccinated for H1N1 is imbalanced.

Taking the correlation (also known as the phi coefficient) between the two target variables a value of approximately $\phi = 0.377$ was obtained. This indicates a positive correlation between the two target variables, meaning that these aren't fully independent from each other.

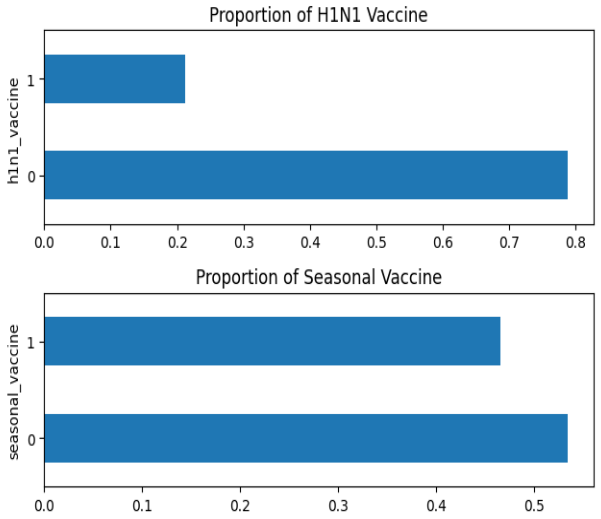

Figure 2.2 - Feature Correlation Heatmap

D. Feature Distributions

Firstly, it is observed that out of the people who received the seasonal vaccine, most of them were female. The same case was also observed with H1N1 vaccine through which one can conclude that women are more prone to get affected than men.

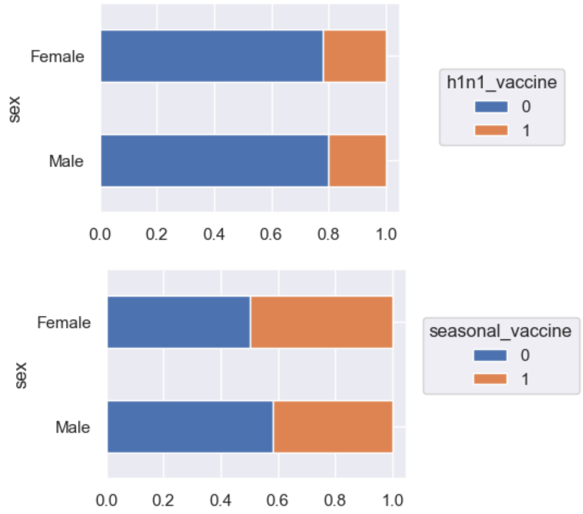

Figure 2 - Vaccination for male and female

The age group has a strong correlation with the seasonal flu vaccine but not with the H1N1 flu vaccine. It seems that people act appropriately when it comes to the seasonal flu as older individuals have a higher risk of complications. However, with H1N1 flu, even though older individuals have a higher risk of complications, they are less likely to get infected. This analysis does not provide information about causality, but it seems that the risk factors are reflected in vaccination rates. It appears that questions related to knowledge and opinions have a strong correlation with both target variables.

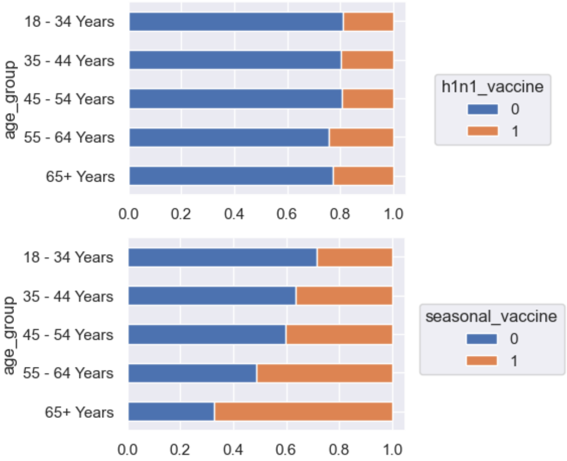

Figure 2.2 - Vaccination for different age groups

III. Performance Metric

To measure the performance of the classifications performed between different classifiers, the ROC (Receiver Operating Characteristic) metric was utilized. The ROC is a type of plot used in binary classifiers, which measures the true positive rate (TPR) against the false positive rate (FPR) for different classifier thresholds. To obtain a quantitative measurement of the performance obtained, it it possible to take the area under the curve (AUC). One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example. Note that this area can be between 0 and 1, the latter characterizing a perfect classifier. Firstly, the AUC was calculated for a *Dummy Classifier*. It is a classifier model that makes predictions without trying to find patterns in the data, serving as a simple baseline to compare with other more complex classifiers. The strategy used to generate predictions was ”uniform”, in order to generate predictions uniformly at random. The value obtained for the AUC of the H1N1 vaccine can be seen in the graph in figure 3.1.

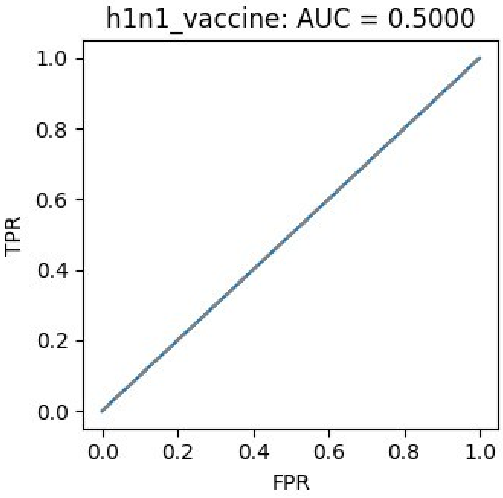

Figure 3.1 - ROC curve for *Dummy Classifier* (H1N1 vaccine)

After that, using the same method, the classification performance of different experimented classifiers was measured. One of the best results was obtained with the *CatBoost* model, whose graph for the seasonal vaccine is represented in figure 3.2.

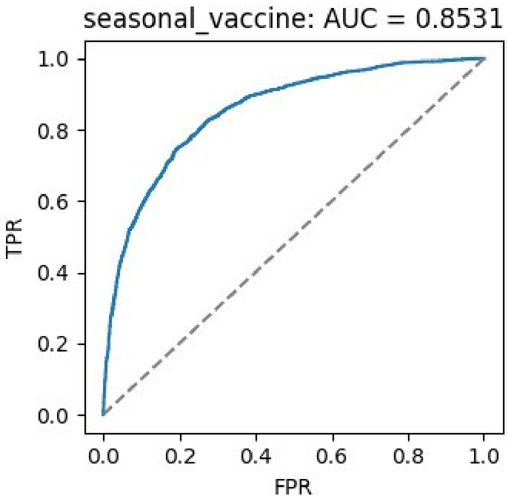

Figure 3.2 - ROC curve for *CatBoost* (seasonal vaccine)

IV. Methodology

As mentioned previously, several classification models were experimented with and compared. Some of the standard classification models utilized for the purposes of this project included Logistic Regression, Multinomial Naive Bayes, K-Nearest neighbors and Decision Trees. Besides these, two gradient boosting algorithms, *CatBoost* and *XGBoost* were also used. For most of the models, in order to automate the data processing and estimation steps, *SKLearn* pipelines were utilized. Within the preprocessing steps, column transformations were performed separately for the numerical and categorical features. For the numerical processing, the `StandardScaler` *SKLearn* class was utilized in most cases to guarantee that each column had nil mean and unit variance, followed by `SimpleImputer` to fill all missing data values utilizing each columns median value.

For categorical features, missing data was also filled using `SimpleImputer()`, but this time with the most frequent value, before being encoded utilizing an One-Hot Encoder, which separated each possible category within each feature into separate binary variables. For the estimation, since most classifiers utilized did not support multi-label classification, the `MultiOutputClassifier()` function was utilized, which in the this case will train two separate instances of the desired estimator.

For measuring the performance of the classifiers obtained before submitting any results for the competition (only 3 daily submissions were allowed on this competition), the training set was split/folded randomly using `train_test_split()` to obtain a performance measurement. After this, the models were trained on the entire dataset. Figure 4.1. illustrates this entire process.

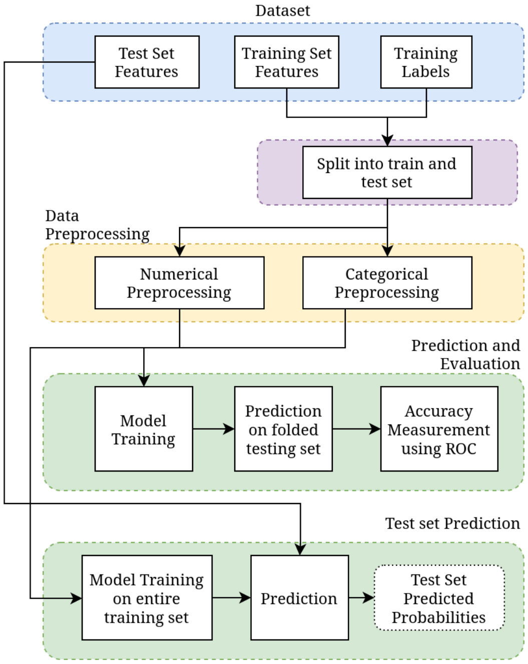
Dataset
Test Set Features
Training Set Features
Training Labels
Split into train and test set
Data Preprocessing
Numerical Preprocessing
Categorical Preprocessing
Prediction and Evaluation
Model Training
Prediction on folded testing set
Accuracy Measurement using ROC
Test set Prediction
Model Training on entire training set
Prediction
Test Set Predicted Probabilities

Figure 4.1 - Model Structure used

A. Logistic Regression

The first model to be trained was the Logistic Regression. Unlike what the name suggests, the Logistic model is a classification model rather than a regression model and is one of the most widely used techniques when it comes to solving classification problems. To utilize the Logistic Regression within our model, *SKLearn* already includes an implementation of this model with the class `LogisticRegression`, which includes several tweakable parameters. To iterate and compare different parameter combinations, *SKLearn* includes a simple way to automate the process with the `GridSearchCV` class. This was utilized for many of the models tested, and in the case of the Logistic Regression, it was utilized to figure out the optimum type of regularization and it's strength.

B. Naive Bayes

The Naive Bayes classification methods, work by applying Bayes' Theorem naively assuming that all the features are conditionally independent from each other. The prior and posterior can then be estimated in different ways.

SKLearn includes several Naive Bayes based classifiers, of which the `GaussianNB` and `MultinomialNB` were utilized in this project. `GaussianNB` assumes that the likelihood of each feature given the label is a gaussian function. Meanwhile, `MultinomialNB` assumes that the data is multinomially distributed, and as such it estimates the posterior probabilities by counting how often a given value appears within each label, for each feature. While using the multinomial model, it was required to remove the standardization/scaling that was applied to the numerical features.

C. K-Nearest Neighbors

The K-Nearest Neighbors is a machine learning algorithm that can be used for both regression and classification. In the case of classification, the algorithm computes the k closest neighbors to the observation, and classifies the observation based on what class held the majority among those neighbors. This method can achieve very good results depending on what value of k is utilized. A larger k will generally suppress the effect of noise on the data, but makes the decision boundary less clear which might affect the ROC score. To use the model, *SKLearn* includes the `KNeighborsClassifier` class, which was used in combination with `GridSearchCV` to find the value k that produces the best results.

D. Decision Trees

Decision Trees are models that attempt to match the target variable by learning decisions from the training data and applying these decisions to the features of an observation's features. These models have the advantage of being conceptually easy to understand, but are quite prone to overfitting the training data.

Similarly to the previous models, the *SKLearn* implementation of the decisions trees, `DecisionTreeClassifier`, was utilized.

E. XGBoost

XGBoost (which stands for eXtreme Gradient Boosting), is a library that provides many gradient boosting algorithms. Essentially, gradient boosting gives a result based on an ensemble of many weaker learning models, as for example decision trees, combining these sequentially and having each stage correct the errors of the previous one.

This library provides many different models and several interfaces for different programming environments. For the purposes of this work, the `XGBClassifier` class was used, as it easily integrates with the *SKLearn*.

F. CatBoost

The final model that was utilized in this project was *CatBoost*. Similarly to XGBoost, and as the name implies, *CatBoost* is also a gradient boosting library. The advantage of *CatBoost* is that it natively supports categorical features without any need for encoding these. This library provides a classification model, also compatible with *SKLearn*, through the class `CatBoostClassifier`.

This model was utilized in two ways, firstly by integrating it into the pipeline developed previously (in the notebook `Project_mainModel.ipynb`, including categorical feature encoding) and in a separate preprocessing pipeline in the file `Project_catBoostedModel.ipynb`. This was done since the first the pipeline utilized *SKLearn's* ColumnTransformers, which made it challenging to specify what columns were categorical in *CatBoost*.

V. Results

Examining the experimental results for all the non gradient boosting models, the following results were obtained:

Model	ROC AUC score
Logistic regression (C = 0.01)	0.84555
Logistic regression (C = 0.1)	0.84655
Logistic regression (C = 1)	0.84640
Logistic regression (C = 10)	0.84635
Gaussian Naive Bayes	0.72947
Multinomial Naive Bayes	0.79467
Decision Tree	0.66738
KNearestNeighbors (k=169)	0.82478

It is possible to see that all the Logistic Regression models exhibited the best classification performance, having pretty much the same score for all the values of C. This parameter is essentially the inverse of the regularization strength. The worst performing model out of all, was clearly the Decision tree classifier. Observing the ROC plot for the Decision Tree, shown in figure 5.1, it is possible to see that it is composed of two straight lines, due to this classifier not being able to estimate the probability of each label, but instead assigning an hard label to each observation.

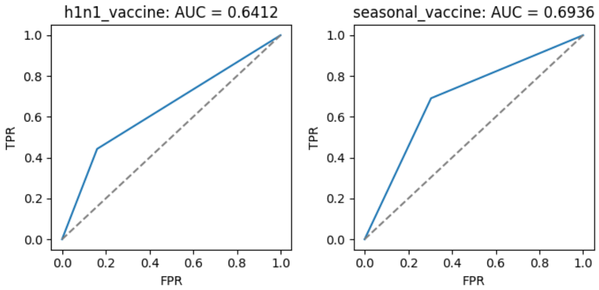

Figure 5.1 - ROC plot for the Decision Tree classifier

In terms of the Naive Bayes classifiers, it is possible to see that the multinomial distribution approximation presents much promising results than the Gaussian approximation.

Taking a look at the results K-Nearest Neighbors, first the score plot of Figure 5.2 was obtained by varying the number of neighbors used and averaging the score across 10 folds of the dataset. It was possible to observe that above a certain threshold of k, that the score seems to stagnate.

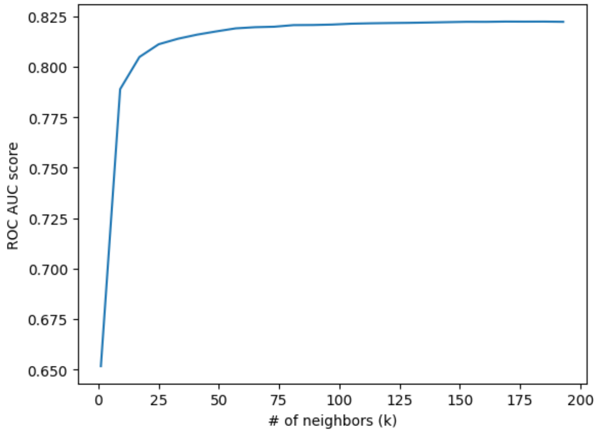

Figure 5.2 - K-Nearest Neighbors ROC AUC Score for varying number of neighbors

VI. Conclusion

VII. Reference Examples

- *Basic format for books:*
J. K. Author, "Title of chapter in the book," in *Title of His Published Book, x*th ed. City of Publisher, (only U.S. State), Country: Abbrev. of Publisher, year, ch. x, sec. x, pp. *xxx–xxx.*
See [1], [2].
- *Basic format for periodicals:*
J. K. Author, "Name of brief," *Abbrev. Title of Periodical*, vol. *x, no.* x,pp. *xxx–xxx,* Abbrev. Month, year, DOI. 10.1109.*XXX*.123456.
See [3]– [5].
- *Basic format for reports:*
J. K. Author, "Title of report," Abbrev. Name of Co., City of Co., Abbrev. State, Country, Rep. *xxx*, year.
See [6], [7].
- *Basic format for handbooks:*
Name of Manual/Handbook, x ed., Abbrev. Name of Co., City of Co., Abbrev. State, Country, year, pp. *xxx–xxx.*
See [8], [9].
- *Basic format for books (when available online):*
J. K. Author, "Title of chapter in the book," in *Title of Published Book*, xth ed. City of Publisher, State, Country: Abbrev. of Publisher, year, ch. x, sec. x, pp. *xxx–xxx.* [Online]. Available: http://www.web.com
See [10]– [13].
- *Basic format for journals (when available online):*
J. K. Author, "Name of brief," *Abbrev. Title of Periodical*, vol. x, no. x, pp. *xxx–xxx*, Abbrev. Month, year. Accessed on: Month, Day, year, DOI: 10.1109.*XXX*.123456, [Online].
See [14]– [16].
- *Basic format for briefs presented at conferences (when available online):*
J.K. Author. (year, month). Title. presented at abbrev. conference title. [Type of Medium]. Available: site/path/file
See [17].
- *Basic format for reports and handbooks (when available online):*
J. K. Author. "Title of report," Company. City, State, Country. Rep. no., (optional: vol./issue), Date. [Online] Available: site/path/file
See [18], [19].
- *Basic format for computer programs and electronic documents (when available online):*
Legislative body. Number of Congress, Session. (year, month day). *Number of bill or resolution*, *Title*. [Type of medium]. Available: site/path/file
***NOTE:* ISO recommends that capitalization follow the accepted practice for the language or script in which the information is given.**
See [20].
- *Basic format for patents (when available online):*
Name of the invention, by inventor's name. (year, month day). Patent Number [Type of medium]. Available: site/path/file
See [21].
- *Basic formatfor conference proceedings (published):*
J. K. Author, "Title of brief," in *Abbreviated Name of Conf.*, City of Conf., Abbrev. State (if given), Country, year, pp. *xxxxxx.*
See [22].
- *Example for briefs presented at conferences (unpublished):*
See [23].
- *Basic format for patents*:
J. K. Author, "Title of patent," U.S. Patent *x xxx xxx*, Abbrev. Month, day, year.
See [24].
- *Basic format for theses (M.S.) and dissertations (Ph.D.):*
  1) J. K. Author, "Title of thesis," M.S. thesis, Abbrev. Dept., Abbrev. Univ., City of Univ., Abbrev. State, year.
  2) J. K. Author, "Title of dissertation," Ph.D. dissertation, Abbrev. Dept., Abbrev. Univ., City of Univ., Abbrev. State, year.

  See [25], [26].
- *Basic format for the most common types of unpublished references:*
  1) J. K. Author, private communication, Abbrev. Month, year.
  2) J. K. Author, "Title of brief," unpublished.
  3) J. K. Author, "Title of brief," to be published.

  See [27]– [29].
- *Basic formats for standards:*
  1) *Title of Standard*, Standard number, date.
  2) *Title of Standard*, Standard number, Corporate author, location, date.

  See [30], [31].
- *Article number in reference examples:*
See [32], [33].
- *Example when using et al.:*
See [34].

References

[1] G. O. Young, "Synthetic structure of industrial plastics," in *Plastics,* 2nd ed., vol. 3, J. Peters, Ed. New York, NY, USA: McGraw-Hill, 1964, pp. 15–64.

[2] W.-K. Chen, *Linear Networks and Systems.* Belmont, CA, USA: Wadsworth, 1993, pp. 123–135.

[3] J. U. Duncombe, "Infrared navigation—Part I: An assessment of feasibility," *IEEE Trans. Electron Devices*, vol. ED-11, no. 1, pp. 34–39, Jan. 1959, 10.1109/TED.2016.2628402.

[4] E. P. Wigner, "Theory of traveling-wave optical laser," *Phys. Rev*., vol. 134, pp. A635–A646, Dec. 1965.

[5] E. H. Miller, "A note on reflector arrays," *IEEE Trans. Antennas Propagat*., to be published.

[6] E. E. Reber, R. L. Michell, and C. J. Carter, "Oxygen absorption in the earth's atmosphere," Aerospace Corp., Los Angeles, CA, USA, Tech. Rep. TR-0200 (4230-46)-3, Nov. 1988.

[7] J. H. Davis and J. R. Cogdell, "Calibration program for the 16-foot antenna," Elect. Eng. Res. Lab., Univ. Texas, Austin, TX, USA, Tech. Memo. NGL-006-69-3, Nov. 15, 1987.

[8] *Transmission Systems for Communications*, 3rd ed., Western Electric Co., Winston-Salem, NC, USA, 1985, pp. 44–60.

[9] *Motorola Semiconductor Data Manual*, Motorola Semiconductor Products Inc., Phoenix, AZ, USA, 1989.

[10] G. O. Young, “Synthetic structure of industrial plastics,” in Plastics, vol. 3, Polymers of Hexadromicon, J. Peters, Ed., 2nd ed. New York, NY, USA: McGraw-Hill, 1964, pp. 15-64. [Online]. Available: http://www.bookref.com.

[11] *The Founders’ Constitution*, Philip B. Kurland and Ralph Lerner, eds., Chicago, IL, USA: Univ. Chicago Press, 1987. [Online]. Available: http://press-pubs.uchicago.edu/founders/

[12] The Terahertz Wave eBook. ZOmega Terahertz Corp., 2014. [Online]. Available: http://dl.z-thz.com/eBook/zomega_ebook_pdf_1206_sr.pdf. Accessed on: May 19, 2014.

[13] Philip B. Kurland and Ralph Lerner, eds., *The Founders’ Constitution.* Chicago, IL, USA: Univ. of Chicago Press, 1987, Accessed on: Feb. 28, 2010, [Online] Available: http://press-pubs.uchicago.edu/founders/

[14] J. S. Turner, “New directions in communications,” *IEEE J. Sel. Areas Commun*., vol. 13, no. 1, pp. 11-23, Jan. 1995.

[15] W. P. Risk, G. S. Kino, and H. J. Shaw, “Fiber-optic frequency shifter using a surface acoustic wave incident at an oblique angle,” *Opt. Lett.*, vol. 11, no. 2, pp. 115–117, Feb. 1986.

[16] P. Kopyt *et al., “*Electric properties of graphene-based conductive layers from DC up to terahertz range,” *IEEE THz Sci. Technol.,* to be published. DOI: 10.1109/TTHZ.2016.2544142.

[17] PROCESS Corporation, Boston, MA, USA. Intranets: Internet technologies deployed behind the firewall for corporate productivity. Presented at INET96 Annual Meeting. [Online]. Available: http://home.process.com/Intranets/wp2.htp

[18] R. J. Hijmans and J. van Etten, “Raster: Geographic analysis and modeling with raster data,” R Package Version 2.0-12, Jan. 12, 2012. [Online]. Available: http://CRAN.R-project.org/package=raster

[19] Teralyzer. Lytera UG, Kirchhain, Germany [Online]. Available: http://www.lytera.de/Terahertz_THz_Spectroscopy.php?id=home, Accessed on: Jun. 5, 2014

[20] U.S. House. 102nd Congress, 1st Session. (1991, Jan. 11). *H. Con. Res. 1, Sense of the Congress on Approval of Military Action*. [Online]. Available: LEXIS Library: GENFED File: BILLS

[21] Musical toothbrush with mirror, by L.M.R. Brooks. (1992, May 19). Patent D 326 189 [Online]. Available: NEXIS Library: LEXPAT File: DES

[22] D. B. Payne and J. R. Stern, “Wavelength-switched pas- sively coupled single-mode optical network,” in *Proc. IOOC-ECOC,* Boston, MA, USA, 1985, pp. 585–590.

[23] D. Ebehard and E. Voges, “Digital single sideband detection for interferometric sensors,” presented at the *2nd Int. Conf. Optical Fiber Sensors,* Stuttgart, Germany, Jan. 2-5, 1984.

[24] G. Brandli and M. Dick, “Alternating current fed power supply,” U.S. Patent 4 084 217, Nov. 4, 1978.

[25] J. O. Williams, “Narrow-band analyzer,” Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, USA, 1993.

[26] N. Kawasaki, “Parametric study of thermal and chemical nonequilibrium nozzle flow,” M.S. thesis, Dept. Electron. Eng., Osaka Univ., Osaka, Japan, 1993.

[27] A. Harrison, private communication, May 1995.

[28] B. Smith, “An approach to graphs of linear forms,” unpublished.

[29] A. Brahms, “Representation error for real numbers in binary computer arithmetic,” IEEE Computer Group Repository, Brief R-67-85.

[30] IEEE Criteria for Class IE Electric Systems, IEEE Standard 308, 1969.

[31] Letter Symbols for Quantities, ANSI Standard Y10.5-1968.

[32] R. Fardel, M. Nagel, F. Nuesch, T. Lippert, and A. Wokaun, “Fabrication of organic light emitting diode pixels by laser-assisted forward transfer,” *Appl. Phys. Lett.*, vol. 91, no. 6, Aug. 2007, Art. no. 061103.

[33] J. Zhang and N. Tansu, “Optical gain and laser characteristics of InGaN quantum wells on ternary InGaN substrates,” *IEEE Photon. J.*, vol. 5, no. 2, Apr. 2013, Art. no. 2600111

[34] S. Azodolmolky *et al.*, Experimental demonstration of an impairment aware network planning and operation tool for transparent/translucent optical networks,” *J. Lightw. Technol.*, vol. 29, no. 4, pp. 439–448, Sep. 2011.