2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-1179-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/EMBC46164.2021.9629537

Feature Analysis and Hierarchical Classification of Anxiety Severity during early COVID-19

Binh Nguyen[1,*], Michael Nigro[1], Alice Rueda[1], Sharadha Kolappan[2], Venkat Bhat[2,3], and Sridhar Krishnan[1]

[1]Department of Electrical, Computer, and Biomedical Engineering, Ryerson Univerity, Toronto, ON M5B 2K3
[2]Interventional Psychiatry Program, St. Michael's Hospital, Toronto M5B 1W8
[3]Department of Psychiatry, University of Toronto, Toronto, ON M5S 1A1
[*]Corresponding author: Binh Nguyen, binh.nguyen@ryerson.ca

***Abstract*— Distress, confusion, and anger are common responses to COVID-19. Statistics Canada created the Canadian Perspectives Survey Series (CPSS) to understand social issues and effects of COVID-19 on the Canadian labour force (LF). The evaluation of the health and health-related behaviours were done through surveys collected between April and July. Features are composed of 4600 participants and 62 questions, which include the General Anxiety Disorder (GAD)-7 questionnaire. This work proposes the use of CPSS2 survey data characteristics to identify the level of anxiety within the Canadian population during early stages of COVID-19 and is validated with the use of GAD-7 questionnaire. Minimum redundancy maximum relevance (mRMR) is applied to select the top 20 features to represent user anxiety. During classification, decision tree (DT) and support vector machine (SVM) are used to test the separation of anxiety severity. Hierarchical classification was used which separated the anxiety severity labels into different test sets and classified accordingly. We employ SVM for binary classification with 10-fold cross validation to separate the labels of *Minimal* and *Severe* anxiety to achieve an overall accuracy of 94.77±0.05%. After analysis, a subset of the reduced feature set can be represented as pseudo passive (PP) data, which are passive sensors that can augment qualitative data. The accurate classification provides proxy on what gives rise to anxiety, as well as the ability to provide early interventions. Future works can implement passive sensors to augment PP data and further understand why people cope this way.**

I. Introduction

Mental health is among the greatest cross-national inequities with 80% of those affected living in low- and middle- income countries [1]. Mental health can be affected by internal factors such as a person's physical health and genetic predisposition as well as external factors such as financial insecurity, food insecurity, and lifestyle changes [2], [3]. In the past year, there has been an increase in mental health discussions, due to the rise of the COVID-19 pandemic. It is widely regarded that mental health has deteriorated since the start of the pandemic [4], [5], yet, mental health itself is a broad and obscure topic.

The motivation for this work is to identify characteristics from the CPSS [6] data that are indicative of anxiety for the Canadian LF population. The LF is constituted by employed and unemployed population. The employed are persons that have a job or a business and the unemployed are without work and are actively seeking work. CPSS is a dataset from Statistics Canada that evaluates the physical and mental health of the Canadian population during the early stages of COVID-19. Studies have focused on self perceived mental health as labels for classifying anxiety. We hypothesize that successful identification of survey data characteristics will allow for indirect assessment of anxiety. Instead of using the more general self-perceived mental health responses as labels as in [4], [7], we propose to use the more quantified GAD-7 labels to look at anxiety among the public during the pandemic. We employ novel techniques for feature selection and classification of the data according to the GAD-7 severity levels to better understand what contributes to anxiety and how to provide early interventions.

The rest of this paper is organized as the following. Section II presents a literature review of the key related works. Section III discusses the CPSS data in further detail and Section IV presents the methodologies used for feature selection and classification. Finally, the results are presented in Section V with a discussion on the drawn conclusions in Section VI.

II. Related Works

The COVID-19 pandemic is having a significant impact on the socioeconomic conditions of the vast majority of the general public [4]. Statistics Canada has undertaken a series of surveys regarding the early stages of COVID-19 called the CPSS, which is aimed at assessing the impact of the pandemic on the Canadian LF [6]. A few studies have been conducted on CPSS using perceived mental health categories [4], [7]. The perceived mental health labels are *Excellent, Very Good, Good, Fair, and Poor*. The surveys contain a number of questions asking individuals about their impressions of the pandemic from a health and economic standpoint. The perception of mental health is apparent as the survey series asks numerous questions regarding self-perceived mental health and the causalities associated with positive and negative self-assessments. In particular, GAD-7 test questionnaire is one such metric that is validated by the Diagnostic and Statistical Manual of Mental Disorders (DSM) for the rating of anxiety severity [8]–[10]. The representation of GAD-7 is a more quantified measure of the severity of anxiety as illustrated in Fig.1.

Authorized licensed use limited to: Toronto Metropolitan University Library. Downloaded on August 29,2024 at 19:38:56 UTC from IEEE Xplore. Restrictions apply.

Mental illness and health covers a wide spectrum of mental conditions. More and more studies have applied artificial intelligence to help improve definitions of mental illness, to identify biomarkers, and to classify mental health [11]. Large mental health datasets extracted from social media platforms, such as Reddit, can benefit from the more advanced deep learning approaches [12]. However, the available public mental health data are generally on a smaller-scale. Studies for mental health on smaller datasets used statistical models like logistic regression and random forest [13].

Traditionally, perceived mental health have been used as the standard label in mental health studies. Using self-perceived mental health and anxiety symptoms, Polsky and Gilmour compared food insecurity among Canadians during the COVID-19 pandemic [3]. The study used logistic regression with sociodemographic covariates adjustment. Individuals who reported moderate food insecurity had nearly three times higher odds of reporting lower level of mental health and higher level of anxiety. When compared with severe food insecurity, the odd ratio for mental health and anxiety increased to 4 and 7.6, respectively.

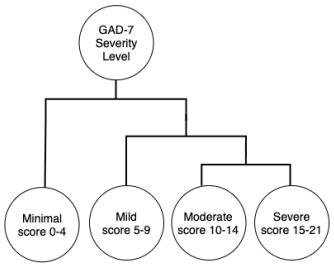


Fig. 1: Hierarchical depiction of the GAD-7 severity levels with cut-off scores.

III. Dataset

The dataset used is from the ongoing data collection effort from Statistics Canada: CPSS [6]. As of the current date, CPSS is composed of four series, which were collected monthly from April to July of 2020. This paper uses CPSS2: Monitoring the Effects of COVID-19. CPSS2 is a set of short surveys that are collected online, at the beginning of COVID-19 from May 4, 2020 until May 10, 2020 to gain information about changes in health and health-related behaviours. The target population are Canadians aged 15 and older in the LF with the exception of full-time members of the Canadian Armed Forces. One individual is randomly selected per household to participate in CPSS.

CPSS2 has 4600 samples points and 62 variables. The variables are grouped to Behaviours and Health impacts (BH), Demographics (DEM), Derived Variables (DV), Food security (FSC), Labour market impacts (LM), Mental health impacts (MH), and Survey related variables (SRV).

CPSS2 contains survey questions, Perceived mental health and GAD-7. This study uses the Severity of GAD-7 as the label, which is derived from GAD-7 questionnaire. GAD-7 is a self diagnostic survey that determines the severity of anxiety disorder. Each question can be scored between 0 to 3, which is composed of 7 questions totalling to a max score of 21. The different severity levels are *Minimal*, *Mild*, *Moderate*, and *Severe anxiety*, which are represented by the score cut-off points of 5, 10, and 15, respectively [8]. Within the target population, 76% and 49.4% of the participants were born in Canada and are male, respectively. While the remaining participants were not born in Canada and are female, respectively. Additionally, Fig. 2 represents the probability distribution of age, household size and marital status of the target population, in respect to the severity of anxiety.

IV. Methods

Prior to applying any analysis, we pre-processed the data around the GAD-7 metric. This involved removing any features directly related to GAD-7 as we use the GAD-7 severity metric as our class label (ANXDVSEV column header). The features related to GAD-7 that are removed are: the 7 questions composed of GAD (MH15A, MH15B, MH15C, MH15D, MH15E, MH15F, MH15G), GAD score (ANXDVGAD), and GAD cut-off (ANXDVGAC). Further pre-processing is done to remove any data samples where no response was provided for the GAD-7 severity metric.

Various feature learning tasks were employed to identify the significant features and mRMR provided the best outcome. mRMR was proposed as feature selection algorithm by optimizing the mutual information (MI) values [14]. The goal was to maximize the distance Φ between the max-Dependency and min-Redundancy as in equation (1). Maximum dependency is computationally expensive, thus, maximum relevance was introduced as a simpler approximation. Maximum relevance (D) between the subset of features $x_i \in S$ and the target class c was obtained as in equation (2). Redundancy among features was estimated by the MI values between two features. Minimum redundancy R calculation is provided in equation (3).

$$\max \Phi(D, R), \Phi = D - R \tag{1}$$

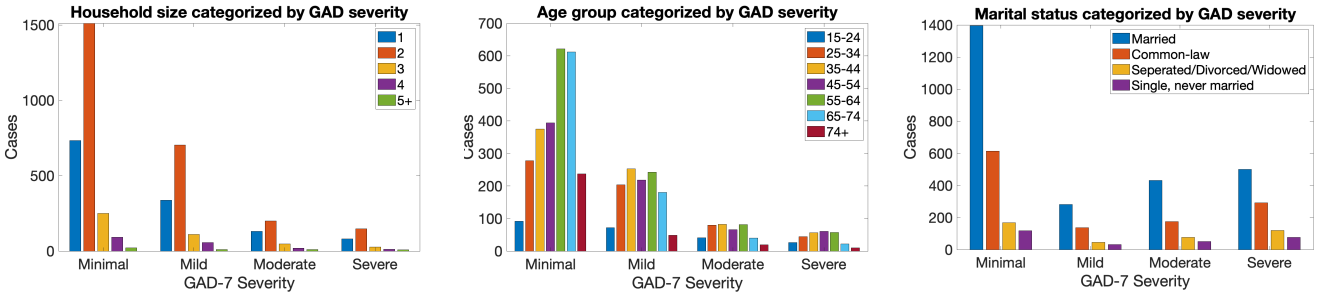
$$\max D(S, c), D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c)) \tag{2}$$

$$\min R(S), R = \frac{1}{|s|^2} \sum_{x_i, x_j \in S} I(x_i; x_j) \tag{3}$$

To validate the selected features, SVM and DT classifiers are used with 10-fold cross-validation to check the veracity of the features using various separations between the labels. SVM and DT classifiers are supervised learning algorithms. Our work uses a linear SVM with one-vs-all approach and binary classification DT. These classifiers were chosen due to the size of our dataset.

A preliminary verification on the selected features was conducted using the greatest distance between the labels, i.e., *Minimal* and *Severe* anxiety. A more granule separation between adjacent labels in a hierarchical grouping were

Authorized licensed use limited to: Toronto Metropolitan University Library. Downloaded on August 29,2024 at 19:38:56 UTC from IEEE Xplore. Restrictions apply.



(a) Household (b) Age (c) Marital Status

Fig. 2: The household, age group and marital statistics of the subjects.

used to further test the robustness of these representative features. We follow GAD-7's hierarchical structure by testing according to the separated branches illustrated in Fig. 1 for the robustness studies.

Precision, recall and F1 score were used as performance metrics for classification on the selected features, as provided in equations (4), (5) and (6), respectively.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

V. Results

After pre-processing 4512 samples remained. mRMR was applied and our work found that 20 features is the minimum number of features required without the sacrifice of classification accuracy of anxiety severity. The reduced feature set can be seen in Table I.

To further support the selection of the reduced feature set, we analyzed the probability distribution of each of the features. Fig. 3 represents BH_35C and BH_40B (Table I). The probability was calculated for each severity level. The probability is equal to the number of sample points per response, divided by the total number of samples in the severity level. Fig. 3b shows tobacco usage and severity of anxiety have a direct correlation. The probability of tobacco usage increased as levels of severity increased. This result supports the findings in [15]. Fig. 3a shows a decline in exercises as severity of anxiety increases. The probability of exercise reduced as severity of anxiety increased as this relation was suggested in [16].

We began our analysis by determining the separation of the labels by focusing on the extremes. We first separated the labels into *Minimal* and *Severe* and classified using 10-fold SVM and DT, achieving an accuracy of 94.77±0.05% and 92.03±0.30%, respectively. The 10-fold SVM approach achieved a recall, precision and F1 score of 98.62±0.09%, 95.72±0.05%, and 97.15±0.05%, respectively. To justify the robustness of our approach, this paper used a hierarchical classification approach where the labels were separated between adjacent labels and tested using an SVM and DT classifier, as shown in Table II.

TABLE I: Reduced feature set through mRMR

Feature	Description	Feature	Description
MH_05	Perceived mental health	LM_40	COVID impact ability meet financial obligations
BH_40D	Eating junk food or sweets	BH_20C	Made plan caring household member are ill
PFSCDV	Household food insecurity	BH_40F	Spending time on the internet
AGEGRP	Age group	SEX	Sex
MHDVMHI	Perceived mental health derived variable	BH_20M	Other precautions taken to reduce risk
BH_20A	Stocked up on essentials	BH_35C	Exercise outdoors
LM35BCDE	EI benefits (sickness/ caregiver/ worksharing/ other)	BH_40A	Consuming alcohol
RURURB	Rural or urban indicator	BH_110	Number of people in close contact
BH_40E	Watching TV	BH_20D	Made a plan for non-household member
BH_35E	Changing food choices	BH_40B	Using tobacco products

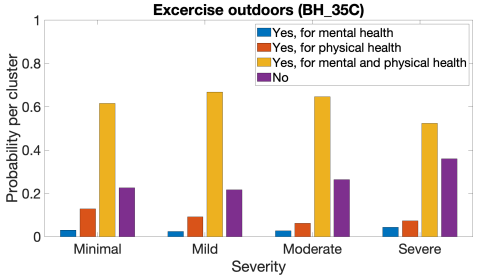
TABLE II: Classification accuracy of hierarchical separation according to adjacent labels

Classes	SVM(%)	DT(%)
Minimal vs Mild, Moderate, and Severe	76.99	68.79
Mild vs Moderate and Severe	71.05	62.64
Moderate vs Severe	63.94	57.52
Overall hierarchical classification average	68.68	62.98

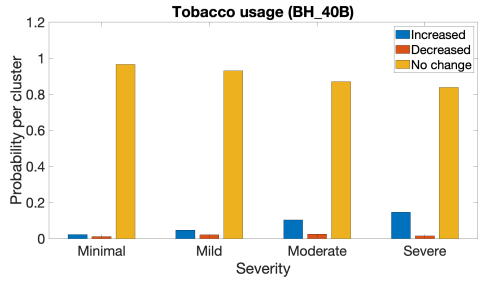
VI. Conclusion and Discussion

The purpose of this work is to analyze the anxiety of the LF during the preliminary stages of COVID-19 using the CPSS2 dataset. This work proposes to use GAD-7 as anxiety severity labels, whereas others used *perceived mental health* [4], [5], [7], [17]. The reason for using GAD-7 is because the scale was developed and validated by DSM, pertaining to

Authorized licensed use limited to: Toronto Metropolitan University Library. Downloaded on August 29,2024 at 19:38:56 UTC from IEEE Xplore. Restrictions apply.



(a) Exercise outdoors



(b) Tobacco Usage

Fig. 3: Probability distribution of responses in respect to severity

more weight and legitimacy to determine one's anxiety. To the author's knowledge, this is the first paper to use GAD-7 for analysis in the CPSS dataset. The application used the reduced 20 feature subset and 10-fold SVM to achieve an overall accuracy of $94.77 \pm 0.05\%$ when comparing the classes *Minimal* and *Severe*. This is a reasonable response as the classes are represented as the opposite extremes of GAD and is further supported by our hierarchical classification results.

Pre-processing and feature selection reduced the number of features used from 62 to 20 features, to improve the efficiency and accuracy of the classifiers. The reduced feature set was determined using mRMR. After analysis of the reduced feature set, it can be determined that many of these features can be augmented as PP data. PP data is qualitative data that can be collected as passive data. For example, within the reduced feature set, BH_35C, BH_40A, BH_40B, BH_40C, BH_40D, BH_40E, BH_40F, BH_110, and RURURB can be coined as PP data (Table I). RURURB can use a GPS to determine a participant's location, BH_35C can use an accelerometer for activity recognition, and BH_40E can determine the audio environment to determine if the participant is watching TV.

The original dataset contained 62 survey questions which can cause survey fatigue, where the participant becomes apathetic or bored, resulting in abandonment of the survey. This work reduced the feature set to 20 while also reducing the potential of survey fatigue. The ability to augment the PP data with a passive sensor in combination with efficient classifiers, give rise to the long-term applications of digital phenotyping. The classification of *Minimal* and *Severe* provides proxy on what gives rise to anxiety, as well as the ability to prepare and provide interventions accordingly. Interventions can be orientated around the features studied in this paper. Future works can incorporate the reduced feature set and augment PP data with passive sensors in their data collection. The study of continuous long-term data collection can further explore and understand how people cope during the COVID-19 pandemic. Techniques related to power requirements, machine learning, and connected healthcare will also be further explored [18], [19].

ACKNOWLEDGMENT

The authors would like to thank NSERC and Mitacs for funding the project.

REFERENCES

[1] K. S. Jacob and V. Patel, "Classification of mental disorders: A global MH perspective," *The Lancet*, vol. 383, pp. 1433–1435, 2014.
[2] L. Martinengo, L. Van Galen, E. Lum, M. Kowalski, M. Subramaniam, and J. Car, "Suicide prevention and depression apps' suicide RA and management: a systematic assessment of adherence to clinical guidelines," *BMC Medicine*, vol. 17, p. 231, dec 2019.
[3] J. Y. Polsky and H. Gilmour, "Food insecurity and MH during the COVID-19 pandemic.," *Health reports*, vol. 31, no. 12, pp. 3–11, 2020.
[4] L. C. Findlay, R. Arim, and D. Kohen, "Understanding the Perceived Mental Health of Canadians During the COVID-19 Pandemic.," *Health reports*, vol. 31, pp. 22–27, jun 2020.
[5] A. Zajacova, A. Jehn, M. Stackhouse, P. Denice, and H. Ramos, "Changes in health behaviours during early COVID-19 and socio-demographic disparities: a cross-sectional analysis," *Canadian Journal of Public Health*, vol. 111, pp. 953–962, dec 2020.
[6] Statistics Canada, "Canadian Perspectives Survey Series 2: Monitoring the effects of COVID-19," 2020.
[7] A. Zajacova, A. Jehn, M. Stackhouse, K. H. Choi, P. Denice, M. Haan, and H. Ramos, "MH and economic concerns from March to May during the COVID-19 pandemic in Canada: Insights from an analysis of repeated cross-sectional surveys," *SSM - Population Health*, vol. 12, p. 100704, Dec 2020.
[8] R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe, "A brief measure for assessing generalized anxiety disorder: the GAD-7.," *Archives of internal medicine*, vol. 166, pp. 1092–7, may 2006.
[9] A. B. Locke, N. Kirst, and C. G. Shultz, "Diagnosis and Management of GAD and Panic Disorder in Adults," Tech. Rep. 9, may 2015.
[10] J. Curtiss and D. H. Klemanski, "Identifying individuals with GAD: A receiver operator characteristic analysis of theoretically relevant measures," *Behaviour Change*, vol. 32, no. 4, pp. 255–272, 2015.
[11] "Artificial Intelligence for Mental Health and Mental Illnesses: an Overview," *Current Psychiatry Reports*, vol. 21, p. 116, nov 2019.
[12] G. Gkotsis, A. Oellrich, S. Velupillai, M. Liakata, T. J. Hubbard, R. J. Dobson, and R. Dutta, "Characterisation of mental health conditions in social media using Informed Deep Learning," *Scientific Reports*, vol. 7, pp. 1–11, 2017.
[13] L. G. Tennenhouse, R. A. Marrie, C. N. Bernstein, and L. M. Lix, "Machine-learning models for depression and anxiety in individuals with immune-mediated inflammatory disease," *Journal of Psychosomatic Research*, vol. 134, p. 110126, jul 2020.
[14] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.
[15] J. L. King, B. A. Reboussin, J. Spangler, J. Cornacchione Ross, and E. L. Sutfin, "Tobacco product use and mental health status among young adults," *Addictive Behaviors*, vol. 77, pp. 67–72, feb 2018.
[16] E. Anderson and G. Shivakumar, "Effects of exercise and physical activity on anxiety," *Frontiers in Psychiatry*, vol. 4, no. APR, 2013.
[17] L.-P. Béland, A. Brodeur, D. Mikola, and T. Wright, "The Short-Term Economic Consequences of COVID-19: Occupation Tasks and Mental Health in Canada," IZA Discussion Papers 13254, Bonn, 2020.
[18] B. Nguyen, Y. Coelho, T. Bastos, and S. Krishnan, "Trends in HAR with focus on machine learning and power requirements," *Machine Learning with Applications*, vol. 5, p. 100072, sep 2021.
[19] S. Krishnan, "Biomedical signal analysis for connected healthcare," *Elsevier*.

Authorized licensed use limited to: Toronto Metropolitan University Library. Downloaded on August 29,2024 at 19:38:56 UTC from IEEE Xplore. Restrictions apply.