Continuous Pain Assessment Using Ensemble Feature Selection from Wearable Sensor Data

Fan Yang[1] , Tanvi Banerjee[1], Mark J. Panaggio[2], Daniel M. Abrams[3], Nirmish R. Shah[4]
[1] Department of Computer Science and Engineering, Wright State University, Dayton, OH, USA
{yang.57, tanvi.banerjee}@wright.edu
[2] Department of Mathematics, Hillsdale College, Hillsdale, MI, USA
mpanaggio@hillsdale.edu
[3] Department of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, IL, USA
dmabrams@northwestern.edu
[4] Division of Hematology, Department of Medicine, Duke University, Durham, NC, USA
nirmish.shah@duke.edu

***Abstract*—Sickle cell disease (SCD) is a red blood cell disorder complicated by lifelong issues with pain. Management of SCD related pain is particularly challenging due to its subjective nature. Hence, the development of an objective automatic pain assessment method is critical to pain management in SCD. In this work, we developed a continuous pain assessment model using physiological and body movement sensor signals collected from a wearable wrist-worn device. Specifically, we implemented ensemble feature selection methods to select robust and stable features extracted from wearable data for better understanding of pain. Our experiments showed that the stability of feature selection methods could be substantially increased by using the ensemble approach. Since different ensemble feature selection methods prefer varying feature subsets for pain estimation, we further utilized stacked generalization to maximize the information usage contained in the selected features from different methods. Using this approach, our best performing model obtained the root-mean-square error of 1.526 and the Pearson correlation of 0.618 for continuous pain assessment. This indicates that subjective pain scores can be estimated using objective wearable sensor data with high precision.**

***Index Terms*—pain assessment, ensemble feature selection, stacked generalization, machine learning**

I. Introduction

Sickle cell disease (SCD) is an inherited red blood cell disorder that can cause a multitude of complications throughout a patient's life. Pain is the most common complication and a significant cause of morbidity. Although pain experienced by SCD patients may become chronic, acute unpredictable vaso-occlusive pain crises lead to frequent visits to the emergency department or day hospital for management [1]. Therefore, an improved understanding of pain, as well as an effective pain management approach is critical. Pain intensity assessment is essential for effective pain management decisions concerning intervention. However, pain is a highly subjective experience, and its assessment is often difficult and relies on self-reports. In clinical practice, medical providers often also consider objective indicators, such as vital signs and non-verbal cues to improve their assessment of pain and create a balance between pain tolerance and medication dosage. Therefore, the development of an objective automatic pain estimation method could lead to improvements in pain assessment and management in SCD.

In recent years, there has been growing interest in developing objective pain assessment techniques based on facial expressions [2]–[4], body movement [5], [6], physiological signals [7]–[9], as well as the fusion of the above data [10], [11]. With the increasing availability of wearable smart devices, it is possible to implement a non-invasive system for health monitoring. Body movement and physiological signals can be easily recorded by wearable devices in real time, which can then be used for automatic pain assessment and improved pain management. In this work, we adopted a wearable wristband (Microsoft Band 2) to be worn by SCD patients in the day hospital settings at Duke University Medical Center as well as University of Pittsburgh Medical Center to record multiple physiological and body movement signals for pain estimation.

The typical steps of the data mining approach for wearable sensors are preprocessing, feature extraction, feature selection and modeling (i.e. learning from data and features to perform tasks such as detection, prediction and decision making) [12]. While numerous features can be extracted from a wearable signal, increasing the number of features does not necessarily increase the model performance since features may be redundant or not indicative of the target variable. Thus, feature selection is used to reduce data dimensionality and eliminate irrelevant and redundant features before machine learning modeling. When applying feature selection in the fields of bioinformatics and biomedicine, both the model performance and the robustness of selected features are equally important. Stable feature selection methods would allow domain experts to have more confidence in the selected features for subsequent analysis. To better understand pain, we implemented four ensemble feature selection methods to select the most robust and stable feature in pain estimation. The ensemble feature selection help to provide knowledge on the pain phenomenon and also yield a more compact and generalizable model.

All ensemble feature selection methods used in this work are

978-1-7281-1867-3/19/$31.00 ©2019 IEEE

Authorized licensed use limited to: Northwestern University. Downloaded on January 26,2023 at 15:50:28 UTC from IEEE Xplore. Restrictions apply.

embedded methods since they have the advantage of performing feature selection and prediction simultaneously, greatly reducing the computational complexity in an ensemble setting. Assuming that each embedded feature selection algorithm will choose the feature subset that is optimal for itself, the pain estimation performance of varying feature subsets chosen by different feature selection methods were evaluated using corresponding regression models. For example, the feature subset chosen by ensemble Random Forest was evaluated using Random Forest regression. The feature subsets selected by different feature selection algorithms are usually inconsistent. Therefore, we employed the stacked generalization, a method that combines multiple learning models by a meta-learner [13]. In this way, we can maximize the usage of information contained in all selected features by different algorithms.

In the present study, we collected physiological and body movement wearable sensor signals from 29 SCD patients during their visits to the day hospital for acute pain. After applying preprocessing and feature extraction to the raw sensor signals, we implemented four ensemble feature selection algorithms to identify the key features of automatic pain assessment. With distinct feature sets selected by different feature selection methods, the corresponding regression models were then used to predict pain on a continuous scale. Furthermore, stacked generalization was applied to combine the four individual learners and optimize information utilization. Our experiments on feature stability show that the robustness of feature selection methods can be significantly improved by extending them with the ensemble procedure. Additionally, the performance of the stacked model indicated the feasibility of using wearable devices to estimate continuous pain intensity.

II. Materials and Methods

A. Data Collection

Patients with SCD presenting for acute pain crisis to the day hospital were approached to participate in the study. Of all patients involved in the study, 20 patients were from Duke University Medical Center and nine patients were from University of Pittsburgh Medical Center. The study included only a one-time visit of each patient. Patients were provided with a Microsoft Band 2 wristband to record physiological and activity measures. Patients were monitored while in the day hospital until the time of discharge with an average duration of 3.61 hours (SD: +/- 1.96 hours). The Microsoft Band 2 has multiple sensors including heart rate monitor, galvanic skin response sensor, skin temperature sensor, three-axis accelerometer and three-axis gyroscope. Overall, we collected ten wearable sensor signals, as shown in Table I, to analyze pain. These ten signals were chosen partially based on signals readily available on the Microsoft Band as well as prior postulated relationships with pain. Patients in more pain typically experience a higher heart rate and move less frequently in the setting of pain [14], [15]. Heart rate variability (HRV) and galvanic skin response (GSR) have been adapted for pain intensity recognition [7], [16], [17]. Furthermore, previous work by our group has supported the use of temperature as a significant predictor of pain for SCD patients [18].

TABLE I
Physiologic and Body Movement Measurements from Microsoft Band 2

Sensor Measurements	Description
Heart Rate(HR)	The number of heartbeats per minute.
RR Interval (RR)	The time interval between successive heartbeats; the measures of specific changes in RR intervals is called heart rate variability (HRV).
Galvanic Skin Response (GSR)	The measure of continuous variation in the electrical characteristics of the skin, also known as skin conductance response (SCR) or electrodermal activity (EDA).
Skin Temperature (SkinTemp)	The temperature of the surface of the skin.
Acceleration in X direction (AccX)	The rate of change of velocity of an object with respect to time in three axis.
Acceleration in Y direction (AccY)	
Acceleration in Z direction (AccZ)	
Angular velocity in X direction (GyroX)[a]	The velocity of an object rotates or revolves in three axis.
Angular velocity in Y direction (GyroY)	
Angular velocity in Z direction (GyroZ)	
Steps (Steps)	The number of accumulated steps per day.

[a] GyroX not correctly captured for some patients and was excluded in the dataset.

Patients were also provided with the mobile-based Technology Resources to Understand Pain (TRU-Pain) app to record pain scores and other symptoms in conjunction with nursing-documented pain scores. Our group has previously reported the usefulness and validity of the mobile health (mHealth) app for patients with SCD [19], [20]. It allowed patients to use a slider bar to rate their pain from 0 (none) to 10 (worst) using numerical rating scale (NRS), thus the pain scores are continuous. Nursing pain scores (in NRS) were also used in the study to enrich the data set and they were assumed to be similar to patients-reported pain scores in the app. Pain scores were reported irregularly with an average of 5.14 (SD: +/- 2.15) records per patients.

B. Preprocessing

The raw wearable sensor signals were retrieved typically every one second in experiments at Duke University Medical Center and every ten seconds at University of Pittsburgh Medical Center. For consistency, the high frequency sensor data (60 data points per minute) were downsampled to the same frequency of the low frequency data (six data points per minute).

By assuming that the pain scores of SCD patients usually do not change rapidly within a short time period, each pain score was matched with the five-minute-long wearable data segment centered on the recording minute of the pain score

Authorized licensed use limited to: Northwestern University. Downloaded on January 26,2023 at 15:50:28 UTC from IEEE Xplore. Restrictions apply.

to ensure that there is sufficient data to extract features. For example, a pain score was reported at 12:05 p.m., then the pain score was matched with the wearable data segment recorded from 12:03 p.m. to 12:07 p.m. (both endpoints were included). Additionally, pain scores without exact time matching were also matched to the wearable sensor data when the timestamp difference between the two data sources, pain recording time and central wearable data segment time, was less than ten minutes. For example, a pain score was reported at 11:45 a.m., and the wearable sensors started recording at 11:52 a.m. Then the pain score was matched with the wearable data segment recorded from 11:52 a.m. to 11:56 a.m. (both endpoints were included). Using this approach, we obtained 149 matched records containing a five-minute-long wearable data segment and a pain score from mobile apps or nurse documents logged during the same (or approximately the same) time period.

C. Feature Extraction

To transform raw sensor signals listed in Table I to a more suitable data representation format, we applied feature extraction on all ten raw signals. Nine features were extracted for each of the ten signals. Table II provides detailed overviews of all features. The feature extraction yielded up to a total of 90 (10×9) features. These extracted features represented the properties of the original raw signals while reducing the volume of data. To reduce the redundant information, all features that correlated positively or negatively with other features at a level of at least 0.95 were eliminated, and 78 features were left in the feature set.

TABLE II
LIST OF FEATURES EXTRACTED FROM WEARABLE SENSOR SIGNALS.

Feature	**Description**
Mean	Average value of the signal.
Standard Deviation	Amount of variation of the signal.
Mean of Derivative	Average rate of change of the signal.
Root Mean Square (RMS)	Square root of the mean of the squares of a set of values.
Peak to Peak	Difference between the maximum and minimum peak.
Peak to RMS	The ratio of the largest absolute value to the RMS value.
Number of Peaks	Number of local maximums (peaks).
Shannon Entropy	For a given signal S, an orthonormal basis and the corresponding coefficients $\{s_i\}$ can be obtained by applying Wavelet Packet Decomposition. Then the Shannon Entropy is this condition is defined as [21]: $E(S) = -\sum_i s_i^2 log(s_i^2)$
Log Energy Entropy	Similar to the Shannon Entropy, the Log Energy Entropy is then defined as: $E(S) = \sum_i log(s_i^2)$

D. Feature Selection Techniques

There are three types of feature selection techniques: filters, wrappers and the embedded methods [22]. Filters select features regardless of the model, therefore these methods are particularly effective in computation time and robust to overfitting. However, filters often consider the features independently, and do not guarantee a feature set with good performance. Wrappers select the subset of features that yields the best possible performance of a given learning algorithm. However, wrapper methods usually need significant computation time which is not feasible in an ensemble feature selection setting. Embedded methods perform feature selection in the process of training and combine the advantages of both previous methods.

Therefore, we adopted four embedded feature selection methods: Lasso Regression (LASSO), Elastic Net (ENet), Random Forest (RF) and Support Vector Machine (SVM) with recursive feature elimination (SVM-RFE). LASSO is a regression model with an L1 penalty, and ENet is a regression model that linearly combines L1 and L2 penalties. Regularized regression with L1 penalty is able to shrink some of the coefficients to zero, thus the feature is removed from the model. In a RF [23], feature importance of each feature is measured by the mean decrease in node impurity over all trees. Then top features can be selected based on feature importances.SVM-RFE [24] starts with all features and removes k features at a time. At each step, the features are ranked according to their weights in the weight vector of a linear SVM, then the k features with the lowest weights are eliminated. The above procedure is repeated until the desired number of features is reached.

E. Ensemble Feature Selection

Given the relatively small sample size in our study, a feature selection method needs to be applied in order to remove irrelevant or redundant features, as well as to prevent overfitting. More importantly, feature selection helps to identify a subset of relevant features which can be used for the knowledge discovery. However, with small sample size, feature selection methods tend to produce inconsistent feature subsets after each run. To increase the stability of the selected features, we applied the ensemble feature selection methods studied by many researchers [25]–[27]. There are two steps in the ensemble feature selection: (1) creating a set of different feature selectors, each with its own outputs (feature rankings or selected feature subset), and (2) aggregating the results of single features selectors to an ensemble output.

Various methods have been exploited for the generation of different feature selectors, which can mainly be divided into two categorizes: data perturbation and function perturbation. Data perturbation runs a feature selection algorithm with different sample subsets, such as bootstrapping [27] and random subsets [28]. Function perturbation involves applying different feature selectors on the same dataset [29], [30]. In this paper, we made use of the data perturbation, more specifically, the bootstrapping method. It is a well-established statistical method which can control and check the stability of results

Authorized licensed use limited to: Northwestern University. Downloaded on January 26,2023 at 15:50:28 UTC from IEEE Xplore. Restrictions apply.

[31]. Given the training data, 100 bootstrap samples were drawn (with replacement) from the training data. Then, a feature selector was applied to each of these bootstrap samples, and 100 diverse sets of features were obtained.

To aggregate the results generated by different feature selectors, linear combination is a simple and effective approach [26]–[28]. For a feature selector that produces a feature ranking (e.g. RF, SVM-RFE), the aggregated ranking is obtained by summing the ranks over all bootstrap samples. For a feature selector that produces a feature subset (e.g. LASSO, ElasticNet), the aggregated feature importance of a feature is then the number of occurrences of the feature over all bootstrap samples. Given a predefined number of features k, the ensemble feature selection algorithm outputs the top k features based on the aggregated feature importance ranking.

Briefly, given the training data and the predefined number of features k, our ensemble feature selection methods linearly combined the feature selection results performed on 100 bootstrapped samples of the training data, and produced the top k features.

A feature selection algorithm is considered stable if the selected feature sets are consistent from multiple runs of the algorithm with variants (such as bootstrapped samples or random subsets of samples) of the dataset. To assess the stability of feature selection techniques, we adopted the Tanimoto distance [25], also known as Jaccard Index. It measures the amount of overlap between two sets (s and s') of arbitrary cardinality, and is defined as:

$$S(s, s') = 1 - \frac{|s| + |s'| - 2|s \cap s'|}{|s| + |s'| - |s \cap s'|}$$

The Tanimoto distance takes values in [0, 1], where 0 means there is no overlap between the two sets, and 1 means the two sets are identical.

F. Regression Methods and Stacked Generalization

While the stability of the feature selection algorithm is important, we also aimed to find the best performing model for continuous pain estimation. Thus, the feature selection needed to be combined with a regression model to predict the pain score on a continuous scale. An important advantage of the four chosen embedded feature selection algorithms is that they integrate model construction with feature selection. Therefore, the corresponding regression models of the four embedded feature selection methods were used to evaluate the pain estimation performance by assuming that the embedded feature selection algorithm will choose the optimal feature subset for the algorithm itself. More specifically, Lasso regression, Elastic Net, Random Forest and Support Vector Machine were used to create regression models based on the chosen feature sets by ensemble LASSO, ensemble ENet, ensemble RF and ensemble SVM-RFE, respectively.

Lasso regression and Elastic Net are both regularized linear regressions modeling the relationship between the target variable and explanatory variables using linear functions. Lasso regression uses only an L1 penalty while Elastic Net uses both L1 and L2 penalties. Random Forest [23] constructs a large number of decision trees at training time and outputs the mean predictions of individual trees. When applying Support Vector Machine in regression [32], the goal is to find a function that deviates from the training output by a value no greater than a certain distance for each training point, and at the same time, is as flat as possible.

Each of the four ensemble feature selection algorithms has its own chosen feature set. To maximize the usage of information contained in the chosen features from different methods, we further adopted the stacked generalization to integrate multiple models. As mentioned above, stacked generalization (also known as stacking) refers to a method that combines multiple learning models with a meta-learner [13]. The base level models are trained on the training set, then the meta-learner learns from the outputs of base level models to increase the learning power beyond the capacity of each individual base level models. The meta-learner is a linear regression model with ridge regularization, while the base level learners are the four ensemble feature selectors (ensemble LASSO, ensemble ENet, ensemble RF, ensemble SVM-RFE) combined with the corresponding regression models. Each base level learner contains an ensemble feature selector and a corresponding regression model. To avoid overfitting, the stacked model in this study was trained and evaluated via a nested 10-fold cross-validation. The procedures of the stacking process can be described as follows (as illustrated in Fig. 1):

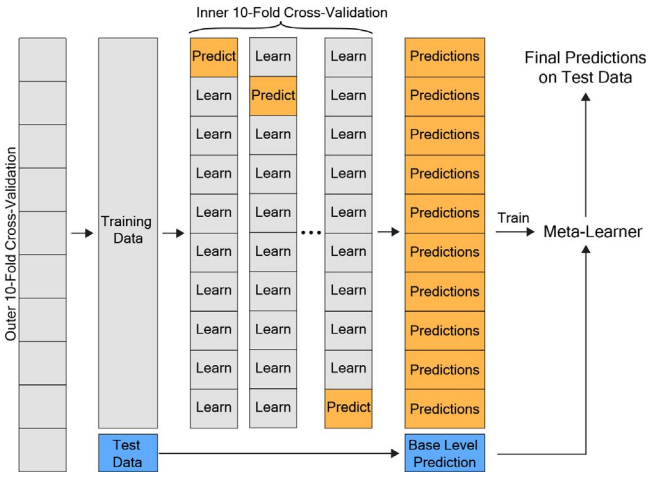

Fig. 1. Illustration of the train and evaluation of the stacked generalization model.

1) Apply the outer 10-fold cross-validation on the entire dataset.
2) At each round of the outer cross-validation, nine folds was the outer training data and a single fold was the outer test data, then the inner 10-fold cross-validation was applied on the outer training data.
3) At each round of the inner cross-validation, each of the four base level model was trained on the inner nine folds and made predictions on the single fold. Then these out-of-folds predictions were combined as the four

Authorized licensed use limited to: Northwestern University. Downloaded on January 26,2023 at 15:50:28 UTC from IEEE Xplore. Restrictions apply.

new features (LASSO predictions, ENet predictions, RF predictions and SVM predictions) for the meta-learner training. Meanwhile, these features were also generated for the outer test set by retraining base level models on the entire outer training data.

4) After the meta-learner was trained on the outer training data, it was evaluated on the outer test set. The final reported performance was averaged among the ten rounds of the outer cross-validation.

III. Results

In this section, two sets of experiments were conducted. The first set of experiments tested the improvement in stability of ensemble feature selection methods by comparing to their single (i.e. non-ensembled) versions. The second set of experiments examined the performance of continuous pain assessment of four base level learners (ensemble feature selection combined with corresponding regression) and the stacked model. Furthermore, the feature importance was analyzed based on the second set of experiments.

A. Stability Results

To estimate the stability of a feature selection algorithm, we used the 10-fold cross-validation. A feature selection algorithm outputted a chosen feature set at each training fold, and ten chosen feature sets were produced in the end. The Tanimoto distance was then computed for each pair of the chosen feature sets. The final stability score of the feature selection algorithm was the average Tanimoto distance over all pairs. A stability score of one means that the ten chosen feature sets are identical. On the other hand, a stability score of zero means that there is no overlap among the ten chosen feature sets. Fig. 2 displays the stability of four feature selection algorithms across different numbers of selected features (a parameter supplied to the ensemble feature selection methods). Each ensemble feature selection method was compared to its single (i.e. non-ensembled) version. In general, it can be observed from Fig. 2 that the ensemble approach, as described in Section II.E, improved the stability as compared to the baseline in all four methods.

Additionally, we calculated the averaged stability (Tanimoto distance) of four feature selection methods over different sizes (as shown in Fig. 2) of chosen feature subsets, ranging from 5 to 78 features. The results are listed in Table III. From Table III, we can observe that the two regularized regression methods (LASSO and ENet) are more stable than RF and SVM-RFE in both single and ensemble versions. On the other hand, the ensemble approach produced more improvement in stability of RF and SVM-RFE than the regularized regression methods. This indicates that less stable algorithms may benefit more

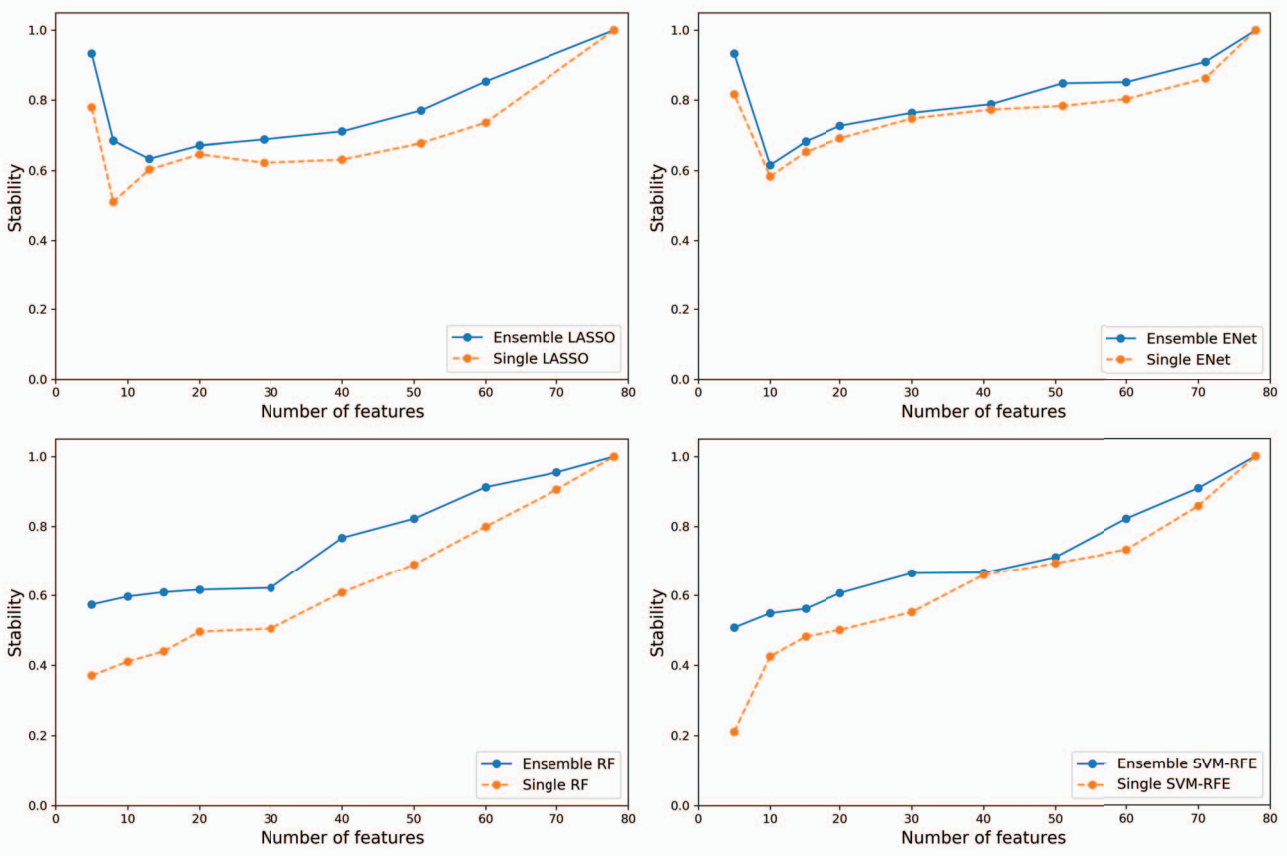

Fig. 2. Stability (Tanimoto distance) of four ensemble feature selection methods and their single versions over different sizes of chosen feature subsets.

Authorized licensed use limited to: Northwestern University. Downloaded on January 26,2023 at 15:50:28 UTC from IEEE Xplore. Restrictions apply.

from the ensemble approach.

TABLE III
AVERAGED STABILITY (TANIMOTO DISTANCE) OF FOUR ENSEMBLE FEATURE SELECTION METHODS AND THEIR SINGLE VERSIONS OVER DIFFERENT SIZES OF CHOSEN FEATURE SUBSETS

	LASSO	ENet	RF	SVM-RFE	Average
Single	0.690	0.771	0.623	0.611	0.674
Ensemble	0.772	0.812	0.748	0.700	0.758
Improvement	0.082	0.041	0.125	0.089	0.084

B. Pain Assessment Results

The pain assessment performance of the four base level models was evaluated by 10-fold cross-validation. A base level model consists of an ensemble feature selection method and the corresponding regression model. At each round of the 10-fold cross-validation, an ensemble feature selection was applied to select a stable feature subset, then the corresponding regression model was built on the selected feature subset. The stacked model combined the four base level models, and was evaluated using the proposed nested 10-fold cross-validation as described in Section II.F. Root Mean Square Error (RMSE) was used as the evaluation metric. RMSE is the square root of the average squared differences between predictions and actual observations. The lower the RMSE, the better the performance of the regression model. Fig. 3 shows the RMSE of the four base level models (LASSO, ENet, RF, SVM-RFE), as well as the stacked model with different numbers of chosen features. The performance of the stacked model is always better than any of the base level models. From Fig. 3, we can also observe that the performances of all four base level models and the stacked model are improved by eliminating irrelevant or redundant features from the full feature sets of 78 features, until the optimal sizes of the feature subsets are reached.

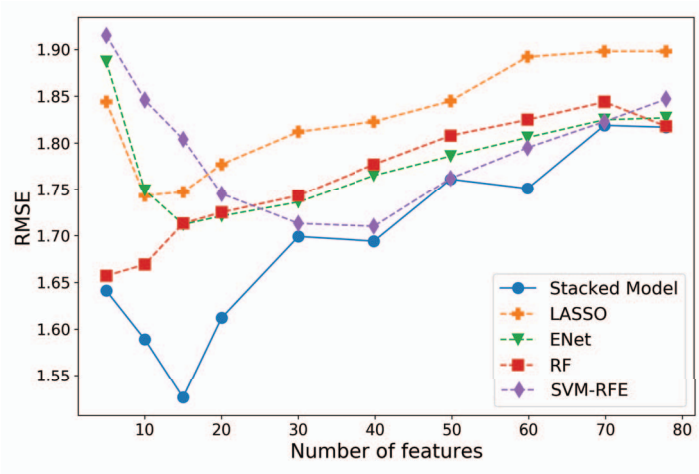

Fig. 3. RMSE of four base level models and the stacked model over varying numbers of chosen features.

The standard deviation of 149 pain scores in the dataset is 1.994, which is equal to the RMSE of a null model that uses mean pain score as a constant prediction. All the regression models in Fig. 3 attained RMSEs lower than this mean-only null model. The best performance of the stacked model was obtained when the selected number of features for each base level models was equal to 15 with the RMSE as 1.526. A Pearson correlation coefficient (linear correlation between predicted values and the actual values) of 0.618 was computed for the best performing model. The strong correlation [33] indicates the feasibility of using wearable sensor signals to predict subjective pain scores with high precision.

C. Feature Importance Analysis

To better understand pain, we further investigated the feature importance in pain estimation. To obtain the feature importance over all four ensemble feature selection methods, we considered a feature as more important if it was selected by more methods. Fig. 4 shows the counts of features selected by the four methods when the predefined number of features was 15. Clearly, different feature selectors preferred different features. In choosing the top 15 of each feature selector, a total of 29 features were selected by all four selectors. These 29 features were used to build the stacked model, which outperformed each single base level model. The stacked model complexity was greatly reduced compared to the full feature sets with 78 features.

According to the source sensor signals, features listed in Fig. 4 can be categorized into six types: (1) heart related features (extracted from HR and RR) (2) galvanic skin response related features (extracted from GSR) (3) skin temperature related features (extracted from SkinTemp) (4) steps related features (extracted from Steps) (5) acceleration related features (extracted from AccX, AccY, AccZ) (6) angular velocity related features (extracted from GyroY, GyroZ). The former three types are physiological measurements while the latter three types are body movement measurements. The type of each feature is indicated by colors in Fig. 4. It can be observed that physiological measurements and body movement measurements are both important in pain assessment. Among the physiological signals, GSR is the most important one followed by HR and RR, and skin temperature is the least important. In heart related features, most of them are related to the variability in heart rate, such as the mean of the derivative of HR (HR_mean_dev) and the standard deviation of RR (RR_std). These results are consistent with many other studies reporting that GSR and heart rate variability (HRV) are significant in pain estimation [17], [34]. For body movement measurements, acceleration and steps are both significant predictors for pain, while angular velocity seems less important but still have made contribution to the pain estimation model. Many features show that body movement is negatively correlated with pain scores. For example, the correlation between the mean number of steps (Steps_mean) and pain is -0.223, and the correlation between the root mean square of AccX (AccX_rms) and pain is -0.258. This may reflect the observation that patients in more pain typically move less frequently [15].

Authorized licensed use limited to: Northwestern University. Downloaded on January 26,2023 at 15:50:28 UTC from IEEE Xplore. Restrictions apply.

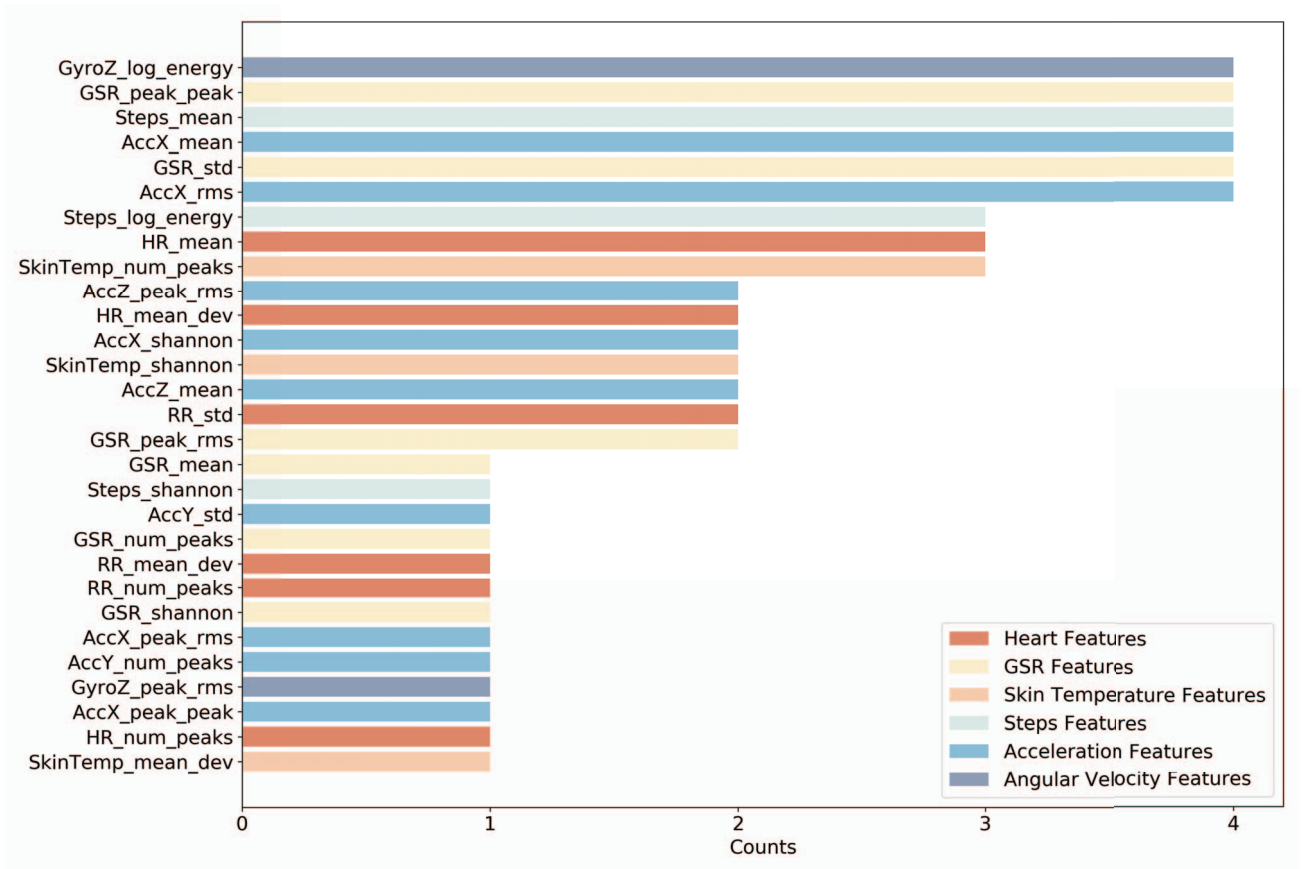

Fig. 4. Feature importance over four ensemble feature selection methods.

D. Conclusion

In this work, we have presented the use of ensemble approach in feature selection. We showed that ensemble feature selection methods considerably increased the robustness and stability of features selected from wearable sensor data. Furthermore, we evaluated the continuous pain estimation performance using each of the four base level learners (ensemble LASSO, ensemble ENet, ensemble RF, ensemble SVM-RFE combined with the corresponding regression models), as well as the stacked model that integrated the four base level learners. The best performance was obtained using the stacked model with RMSE of 1.526 and Pearson correlation of 0.618. We also demonstrated that physiological and body movement measurements were both important in automatic pain estimation using wearable sensors.

Acknowledgment

Research reported in this publication was supported by the National Center For Complementary & Integrative Health of the National Institutes of Health under Award Number R01AT010413. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] D. M. Cline, S. Silva, C. E. Freiermuth, V. Thornton, and P. Tanabe, "Emergency department (ed), ed observation, day hospital, and hospital admissions for adults with sickle cell disease," *Western Journal of Emergency Medicine*, vol. 19, no. 2, p. 311, 2018.

[2] L. Martinez, D. Rosalind Picard *et al.*, "Personalized automatic estimation of self-reported pain intensity from facial expressions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 70–79.

[3] R. Yang, X. Hong, J. Peng, X. Feng, and G. Zhao, "Incorporating high-level and low-level cues for pain intensity estimation," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 3495–3500.

[4] P. Casti, A. Mencattini, M. C. Comes, G. Callari, D. Di Giuseppe, S. Natoli, M. Dauri, E. Daprati, and E. Martinelli, "Calibration of vision-based measurement of pain intensity with multiple expert observers," *IEEE Transactions on Instrumentation and Measurement*, 2019.

[5] J. K. Lee, G. T. Desmoulin, A. H. Khan, and E. J. Park, "A portable inertial sensing-based spinal motion measurement system for low back pain assessment," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 4737–4740.

[6] T. A. Olugbade, N. Bianchi-Berthouze, N. Marquardt, and A. C. Williams, "Pain level recognition using kinematics and muscle activity for physical rehabilitation in chronic pain," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 243–249.

[7] S. Gruss, R. Treister, P. Werner, H. C. Traue, S. Crawcour, A. Andrade, and S. Walter, "Pain intensity recognition rates via biopotential feature patterns with support vector machines," *PloS one*, vol. 10, no. 10, p. e0140330, 2015.

Authorized licensed use limited to: Northwestern University. Downloaded on January 26,2023 at 15:50:28 UTC from IEEE Xplore. Restrictions apply.

[8] Y. Chu, X. Zhao, J. Han, and Y. Su, "Physiological signal-based method for measurement of pain intensity," *Frontiers in neuroscience*, vol. 11, p. 279, 2017.

[9] D. Lopez-Martinez and R. Picard, "Multi-task neural networks for personalized pain recognition from physiological signals," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 2017, pp. 181–184.

[10] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. C. Traue, "Automatic pain recognition from video and biomedical signals," in *2014 22nd International Conference on Pattern Recognition*. IEEE, 2014, pp. 4582–4587.

[11] G. Zamzmi, C.-Y. Pai, D. Goldgof, R. Kasturi, T. Ashmeade, and Y. Sun, "An approach for automated multimodal analysis of infants' pain," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 4148–4153.

[12] H. Banaee, M. Ahmed, and A. Loutfi, "Data mining for wearable sensors in health monitoring systems: a review of recent trends and challenges," *Sensors*, vol. 13, no. 12, pp. 17 472–17 500, 2013.

[13] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[14] Y. Tousignant-Laflamme, P. Rainville, and S. Marchand, "Establishing a link between heart rate and pain in healthy subjects: a gender effect," *The journal of pain*, vol. 6, no. 6, pp. 341–347, 2005.

[15] P. W. Hodges, "Pain and motor control: from the laboratory to rehabilitation," *Journal of Electromyography and Kinesiology*, vol. 21, no. 2, pp. 220–228, 2011.

[16] M. Kächele, M. Amirian, P. Thiam, P. Werner, S. Walter, G. Palm, and F. Schwenker, "Adaptive confidence learning for the personalization of pain intensity estimation systems," *Evolving Systems*, vol. 8, no. 1, pp. 71–83, 2017.

[17] D. Lopez-Martinez and R. Picard, "Continuous pain intensity estimation from autonomic signals with recurrent neural networks," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 5624–5627.

[18] F. Yang, T. Banerjee, K. Narine, and N. Shah, "Improving pain management in patients with sickle cell disease from physiological measures using machine learning techniques," *Smart Health*, vol. 7, pp. 48–59, 2018.

[19] N. Shah, J. Jonassaint, and L. De Castro, "Patients welcome the sickle cell disease mobile application to record symptoms via technology (smart)," *Hemoglobin*, vol. 38, no. 2, pp. 99–103, 2014.

[20] C. R. Jonassaint, N. Shah, J. Jonassaint, and L. De Castro, "Usability and feasibility of an mhealth intervention for monitoring and managing pain symptoms in sickle cell disease: the sickle cell disease mobile application to record symptoms via technology (smart)," *Hemoglobin*, vol. 39, no. 3, pp. 162–168, 2015.

[21] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Transactions on information theory*, vol. 38, no. 2, pp. 713–718, 1992.

[22] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[23] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[24] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[25] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: a study on high-dimensional spaces," *Knowledge and information systems*, vol. 12, no. 1, pp. 95–116, 2007.

[26] Y. Saeys, T. Abeel, and Y. Van de Peer, "Robust feature selection using ensemble feature selection techniques," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2008, pp. 313–325.

[27] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," *Bioinformatics*, vol. 26, no. 3, pp. 392–398, 2009.

[28] N. Meinshausen and P. Bühlmann, "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 4, pp. 417–473, 2010.

[29] J. Dutkowski and A. Gambin, "On consensus biomarker selection," *BMC bioinformatics*, vol. 8, no. 5, p. S5, 2007.

[30] M. Netzer, G. Millonig, M. Osl, B. Pfeifer, S. Praun, J. Villinger, W. Vogel, and C. Baumgartner, "A new ensemble-based algorithm for identifying breath gas marker candidates in liver disease using ion molecule reaction mass spectrometry," *Bioinformatics*, vol. 25, no. 7, pp. 941–947, 2009.

[31] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. CRC press, 1994.

[32] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[33] J. D. Evans, *Straightforward statistics for the behavioral sciences.* Thomson Brooks/Cole Publishing Co, 1996.

[34] H. Lim, B. Kim, G.-J. Noh, and S. K. Yoo, "A deep neural network-based pain classifier using a photoplethysmography signal," *Sensors*, vol. 19, no. 2, p. 384, 2019.

Authorized licensed use limited to: Northwestern University. Downloaded on January 26,2023 at 15:50:28 UTC from IEEE Xplore. Restrictions apply.