**APPENDIX A : LEXICON FOR DEVELOPMENT OF CLASSIFICATION ALGORITHM**

|  |  |  |  |
| --- | --- | --- | --- |
| Outcomes of interest | Base string | Classification Labels | Associated keywords/Phrases/Values |
| Mitral Regurgitation | Mitral Regurgitation, MR | 1. Mild | Trace, possible mitral regurgitation present, possible MS present, trivial, least, clinically insignificant, trace-mild, Mild, minimal, trace/mild, trace-to-mild, no significant, insignificant , minimal, midl |
| 2. Mild to Moderate | Mild to Moderate, mild-to-moderate, mild-mod, mild-tp-moderate, mild-to-mod, mild-tp-moserate, mild-to-moserate, mild-moserate, midl to moderate |
| 3. Moderate | Moderate, modate, mdrate |
| 4. Moderate to Severe | Moderate-to-severe,mod-sev,moderate-tp-severe,mod-severe,moderate but cannot exclude severe, moderate to severe, mod-tp-sev, mod-tp-severe, moderate-tp-sev, moderately severe |
| 5. Severe | Severe, sev, critical |
| 6. Indeterminate | Absence of above Associated keywords/Phrases (1-5), cannot rule out MR, cannot rule out Mitral regurgitation, can’t rule out, cannot exclude |
| 7. Normal Cases (negative) | Absence of the above keywords/phrases and base string, No, no evidence of, uncertain, without |
| Mitral Stenosis | Mitral Stenosis, MS | 1. Mild | Trace, possible mitral stenosis present, MS present, trivial, least, Clinically insignificant, trace-mild, Mild, minimal, trace/mild, trace-to-mild, no significant, insignificant , minimal, midl |
| 2. Mild to Moderate | Mild to Moderate, mild-to-moderate, mild-mod, mild-tp-moderate, mild-to-mod, mild-tp-moserate, mild-to-moserate, mild-moserate, midl to moderate |
| 3. Moderate | Moderate, modate, mdrate |
| 4. Moderate to Severe | Moderate-to-severe,mod-sev,moderate-tp-severe,mod-severe,moderate but cannot exclude severe, moderate to severe, mod-tp-sev, mod-tp-severe, moderate-tp-sev, moderately severe |
| 5. Severe | Severe, sev, critical |
| 6. Indeterminate | Absence of above Associated keywords/Phrases (1-5), cannot rule out MR, cannot rule out Mitral regurgitation, can’t rule out, cannot exclude |
| 7. Normal Cases (negative) | Absence of the above keywords/phrases and base string, No, no evidence of, uncertain, without |
| Aortic Regurgitation | Aortic regurgitation, AR | 1. Mild | Trace, possible mitral stenosis present, MS present, trivial, least, clinically insignificant, trace-mild, mild, minimal, trace/mild, trace-to-mild, no significant, insignificant , minimal, midl |
| 2. Mild to Moderate | mild to moderate, mild-to-moderate, mild-mod, mild-tp-moderate, mild-to-mod, mild-tp-moserate, mild-to-moserate, mild-moserate, midl to moderate |
| 3. Moderate | moderate, modate, mdrate, mod |
| 4. Moderate to Severe | moderate-to-severe,mod-sev,moderate-tp-severe,mod-severe,moderate but cannot exclude severe, moderate to severe, mod-tp-sev, mod-tp-severe, moderate-tp-sev, moderately severe |
| 5. Severe | Severe, sev, critical |
| 6. Indeterminate | Absence of above Associated keywords/Phrases (1-5), cannot rule out MR, cannot rule out Mitral regurgitation, can’t rule out, cannot exclude |
| 7. Normal Cases (negative) | Absence of the above keywords/phrases and base string, No, no evidence of, uncertain, without |
| Aortic Stenosis | Aortic valve stenosis, aortic stenosis, AS | 1. Mild | Trace, possible mitral stenosis present, MS present, trivial, least, Clinically insignificant, trace-mild, mild, minimal, trace/mild, trace-to-mild, no significant, insignificant , minimal, midl |
| 2. Mild to Moderate | mild to moderate, mild-to-moderate, mild-mod, mild-tp-moderate, mild-to-mod, mild-tp-moserate, mild-to-moserate, mild-moserate, midl to moderate |
| 3. Moderate | moderate, modate, mdrate, mod |
| 4. Moderate to Severe | moderate-to-severe,mod-sev,moderate-tp-severe,mod-severe,moderate but cannot exclude severe, moderate to severe, mod-tp-sev, mod-tp-severe, moderate-tp-sev, moderately severe |
| 5. Severe | severe, sev, critical |
| 6. Indeterminate | absence of above associated keywords/Phrases (1-5), cannot rule out MR, cannot rule out mitral regurgitation, can’t rule out, cannot exclude |
| 7. Normal Cases (negative) | absence of the above keywords/phrases and base string, No, no evidence of, uncertain, without |
| Flow Gradient | Aortic valve stenosis/aortic stenosis/AS along with ‘Low flow’, ‘low flow’,’ low-flow’ | 1.Low Flow Low Gradient | When normal and paradoxical is absent |
| 2. Paradoxical Low Flow Low Gradient | paradoxical |
| 3.Normal Flow Gradient | Normal flow or absence of above phrases/keywords and base string |
| Diastolic Function | Diastolic function/Grade | 1.Normal | Normal diastolic function, low-normal diastolic function |
| 2.Grade I | Grade I |
| 3.Grade II | Grade II |
| 4.Grade III | Grade III |
|  |  | 5.Indeterminate | Unable to assess |
| Ejection Fraction | Ejection fraction, LVEF, EF, estimated EF left ventricular ejection, 3D EF, fraction, left ventricular systolic function, systolic ventricle, cardiac function, ventricular contractility, systolic contractility, | 1.Numeric Values | All numeric values |
| 2.Ranges | All ranges |
| 3.Qualitative | Around, severe, moderate, mild to moderate, moderate to severe, poor, sluggish, reduce, decrease, depress, impair, abnormal, preserved, abnormal, intact, adequate, good, satisfactory, excellent, hyperdynamic, hyperkinetic, vigorous, |
|  |  | 4.Not available or missing | Technical difficulty, cannot estimate, unable to assess |
| Mitraclip | Mitral valve | Mitra Clip (Yes=1,N0=1) | mitraclip |
| Prosthetic | Mitral valve | Prosthetic Valve, (Yes=1,N0=1) | Bioprosthetic, prosthetic, paravalvular, transvalvular, mechanical, surgical ring, TAVR |

Table 1: Lexicon consisting of keywords and phrases to define the classification algorithm rule from 2,000 samples of extracted echo semi-structured data

**APPENDIX B: LEXICON FOR ASSIGNMENT OF EJECTION FRACTION FROM KEYWORDS**

|  |  |
| --- | --- |
| Keywords | Ejection fraction assignment |
| Severe | 16 |
| moderate, poor, moderate, sluggish, reduce, decreased, depressed, impair, abnormal, below normal, mild, mild to moderate | 45 |
| normal, no abnormal, intact, preserved, adequate, good, satisfactory, excellent | 55 |
| hyperdynamic, hyperkinetic, vigorous | 70 |

Table 2: Ejection fraction assignment rule for qualitative EF.

**APPENDIX C: FEATURE ENGINEERING DETAILS**

1. **pyodbc:** We used pyodbc package in python 3.6 for connecting to the SQL server to retrieve the echo data stored in our database.
2. **pandas:** We used python 3.6 pandas library for data manipulation and analysis.
3. **re:** We used regular expression in python for text processing [cpython/re.py at 3.10 · python/cpython · GitHub](https://github.com/python/cpython/blob/3.10/Lib/re.py)
4. **nltk:** We used ‘english’ stopwords from nltk library and ‘tokenize’ to achieve tokenization in python 3.6.
5. **sklearn:** We used ‘countvectorizer’ to create a matrix of token counts. max\_df and min\_df has default settings. We ignored terms that appear in more than 100% of the echo reports and less than one echo report. We used ‘fit\_transform’ for scaling and used ‘TfidfTransformer’ to transform a count matrix to a ‘tf-idf’ representation. ‘labelencoder’ is used to encode the target values in python 3.6.
6. **imblearn:** We used imblearn library in python 3.6 and utilized ‘smotenn’ for oversampling and ‘EditedNearestNeighbours’ for under sampling the minority class. The sampling strategy is set to ‘all’

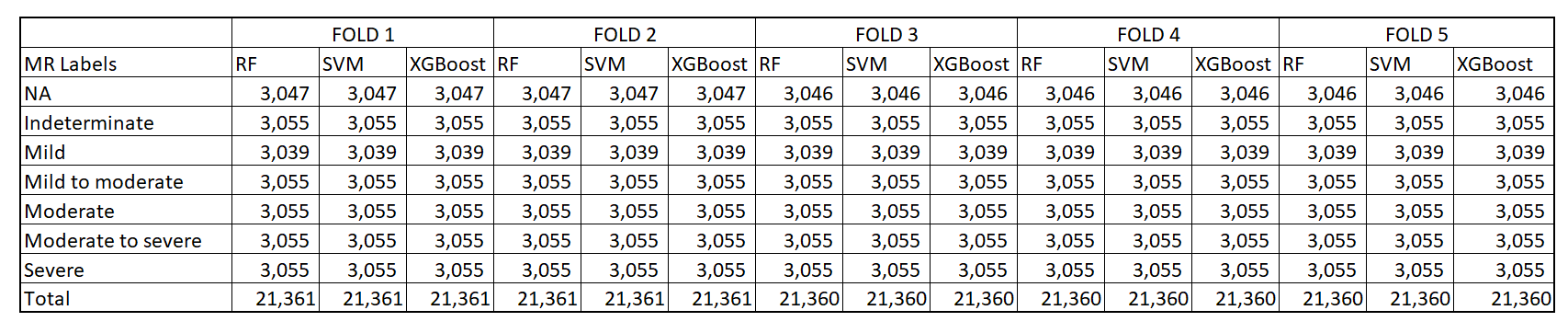
**APPENDIX D: MACHINE LEARNING IMPLEMENTATION DETAILS**

1. **StratifiedKFold:** We used StratifiedKFold obtain training and test data with 5 folds to obtain equal number of training and test sets. The ‘random state’ is set to 42 and ‘shuffle’ is set to ‘true’ in python 3.6.
2. **classification\_report:** ‘classification\_report’ from ‘sklearn.metrics’ is leveraged to define the classification report with accuracy score in python 3.6.
3. **LinearSVC:** ‘linear SVC’ is imported from sklearn.svm to train the SVM model in python 3.6.
4. **RandomForestClassifier: ‘**RandomForestClassifier’ is imported from sklearn. ensemble to train the ‘Random Forest’ model in python 3.6.
5. **XGBClassifier:** ‘XGBClassifier’ is imported from ‘xgboost’ to train the extreme gradient boosting model in python 3.6.

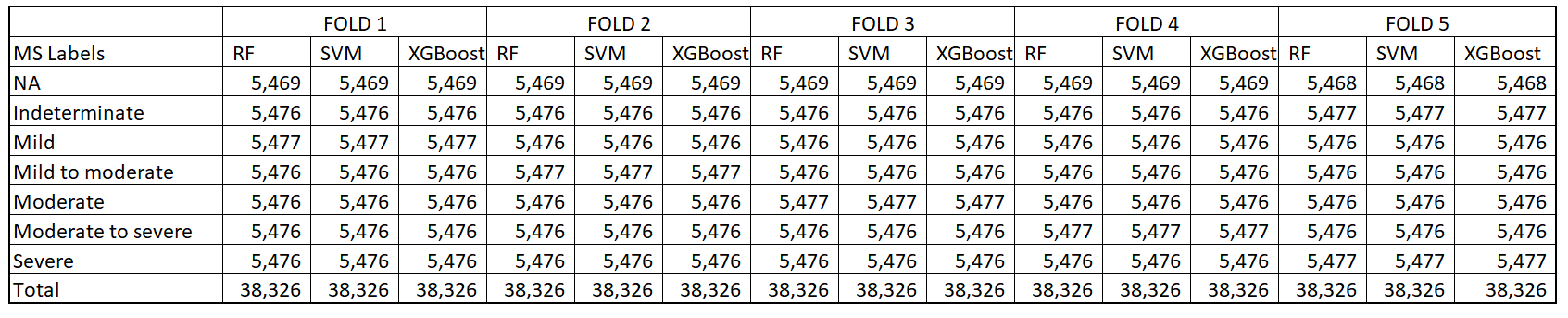
**APPENDIX E: DATA RESAMPLING RESULTS**

|  |  |  |  |
| --- | --- | --- | --- |
| **Outcomes of Interest** | **Stratified Folds** | **Training set size** | **Test set sample size** |
| **Mitral Regurgitation** | Fold 1,2 | 85,441 | 21,361 |
| Fold 3,4,5 | 85,442 | 21,360 |
| **Mitral Stenosis** | Fold 1 to 5 | 153,304 | 38,326 |
| **Aortic Regurgitation** | Fold 1,2,3,4 | 115,347 | 28,837 |
| Fold 5 | 115,348 | 28,836 |
| **Aortic Stenosis** | Fold 1 | 146,372 | 36,594 |
| Fold 2,3,4,5 | 146,373 | 36,593 |
| **Flow Gradient across Aortic Stenosis** | Fold 1,2 | 66,913 | 16,729 |
| Fold 3,4,5 | 66,914 | 16,728 |
| **Diastolic Function** | Fold 1,2 | 39,485 | 9,872 |
| Fold 3,4,5 | 39,486 | 9,871 |
| **Ejection Fraction (Numeric)** | Fold 1 to 5 | 34,572 | 8,643 |
| **Ejection Fraction (Ranges)** | Fold 1,2,3 | 94,218 | 23,555 |
| Fold 4,5 | 94,219 | 23,554 |
| **Ejection Fraction (Qualitative)** | Fold 1,2,3 | 506 | 127 |
| Fold 4,5 | 507 | 126 |

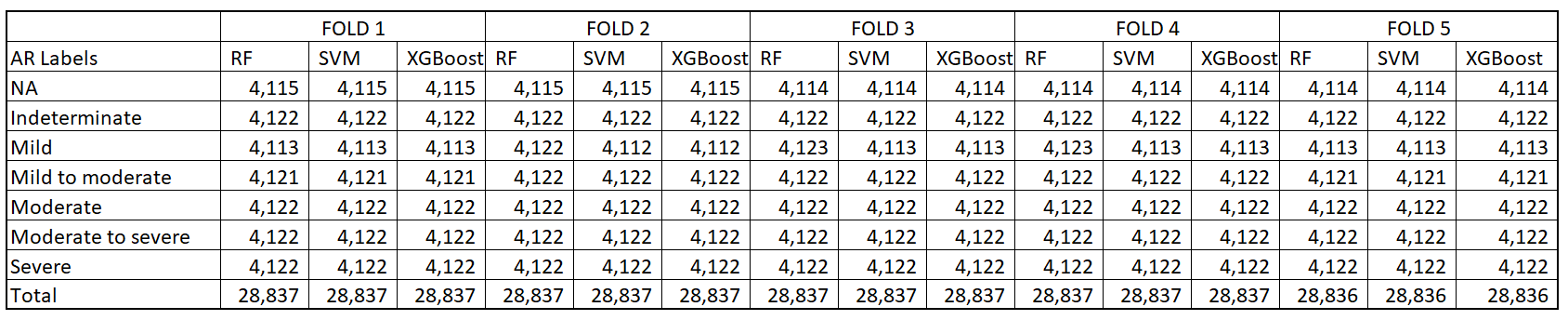
**APPENDIX F(1): STRATIFIED SUPPORT SAMPLES FOR EACH CLASS IN A BALANCED TEST SET OBTAINED USING SMOTE – ENN METHOD IN FIVE FOLDS FOR MITRAL REGURGITATION**

****

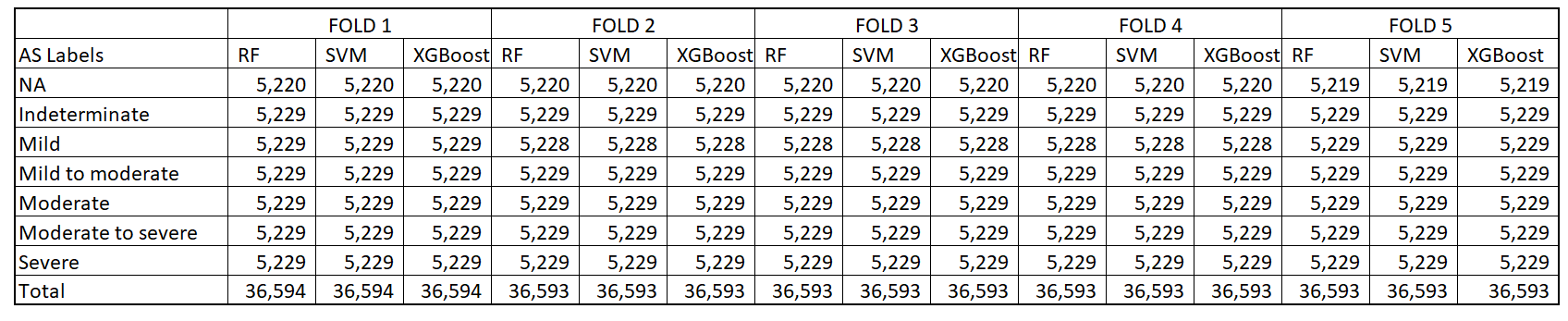
**APPENDIX F(2): STRATIFIED SUPPORT SAMPLES FOR EACH CLASS IN A BALANCED TEST SET OBTAINED USING SMOTE – ENN METHOD IN FIVE FOLDS FOR MITRAL STENOSIS**

****

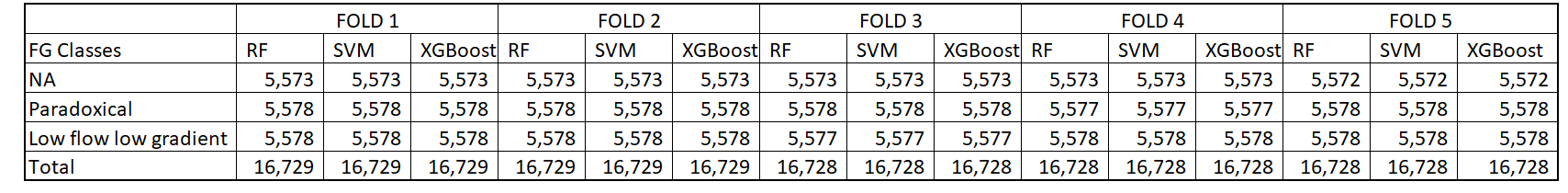
**APPENDIX F(3): STRATIFIED SUPPORT SAMPLES FOR EACH CLASS IN A BALANCED TEST SET OBTAINED USING SMOTE – ENN METHOD IN FIVE FOLDS FOR AORTIC REGURGITATION**

****

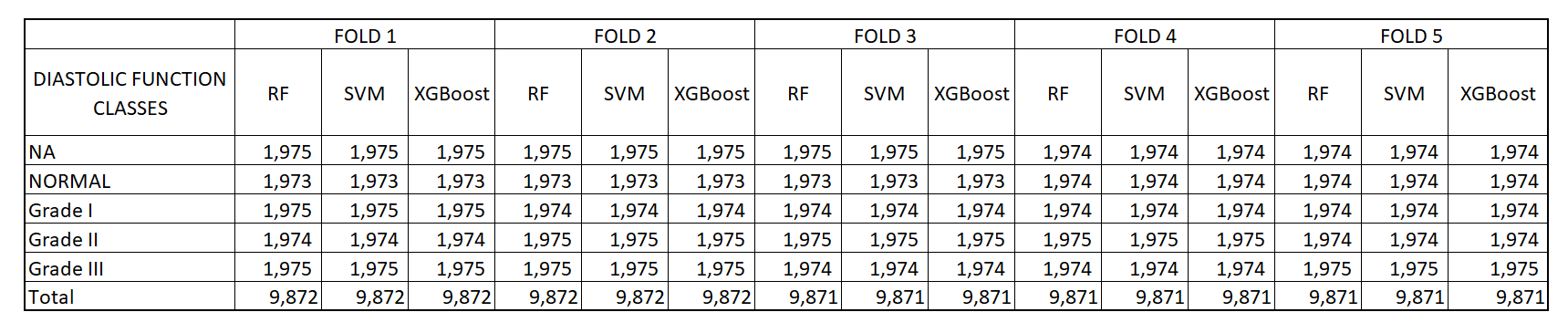
**APPENDIX F(4): STRATIFIED SUPPORT SAMPLES FOR EACH CLASS IN A BALANCED TEST SET OBTAINED USING SMOTE – ENN METHOD IN FIVE FOLDS FOR AORTIC STENOSIS**

****

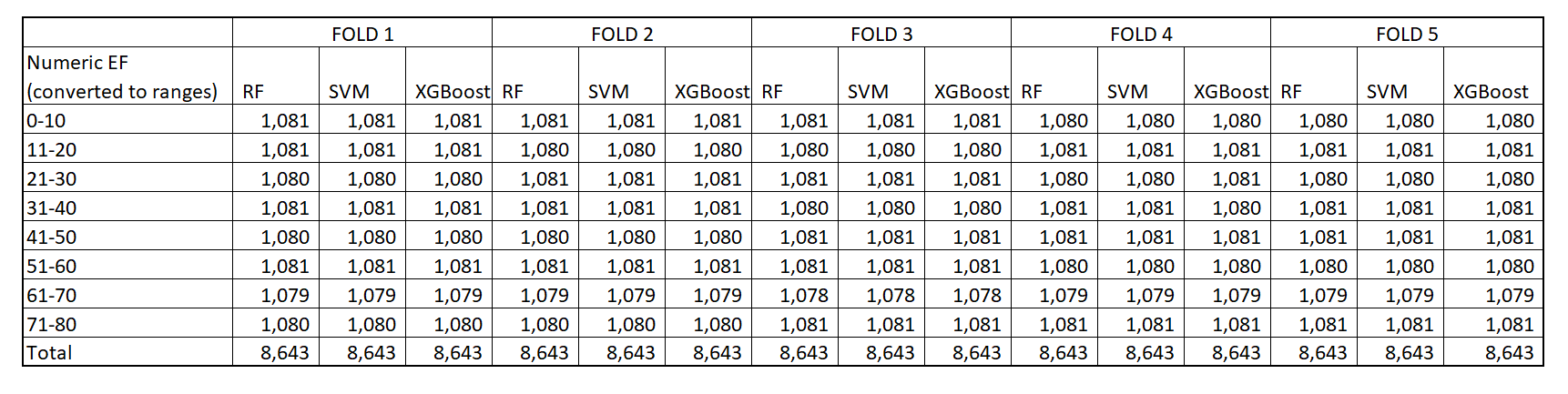
**APPENDIX F(5): STRATIFIED SUPPORT SAMPLES FOR EACH CLASS IN A BALANCED TEST SET OBTAINED USING SMOTE – ENN METHOD IN FIVE FOLDS FOR FLOW GRADIENT OF AORTIC STENOSIS**

****

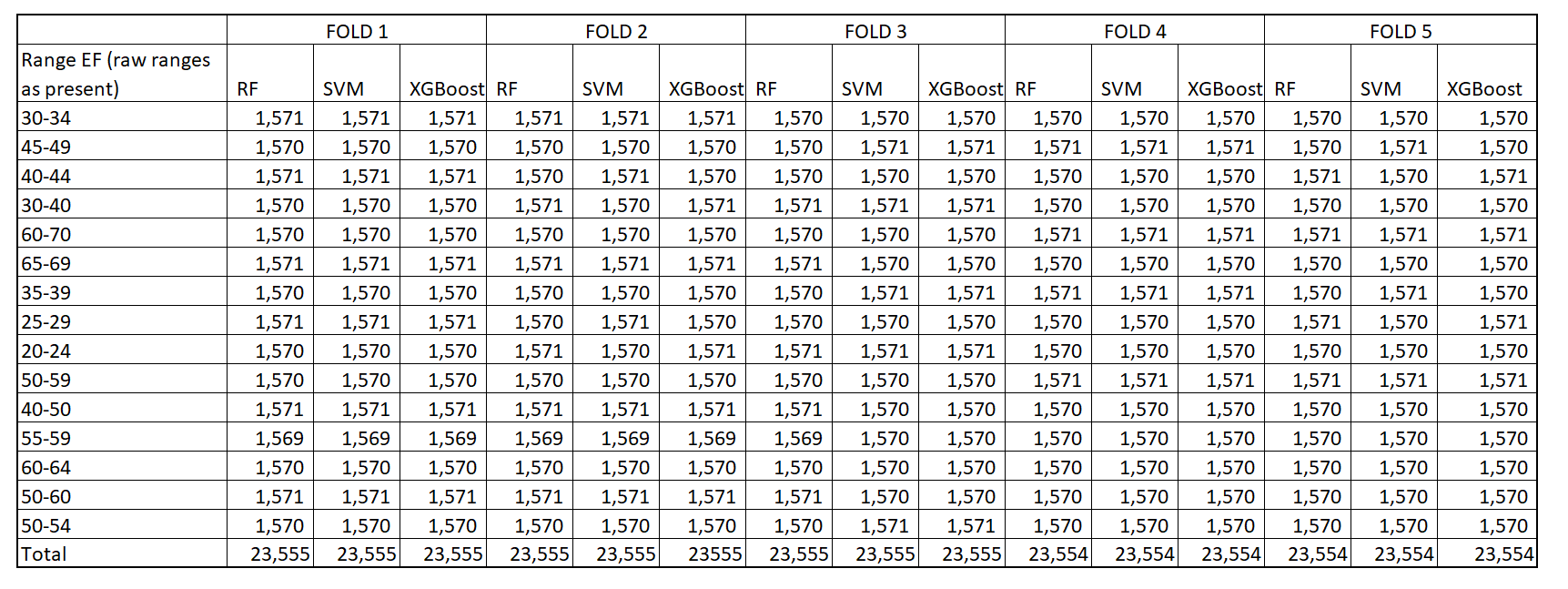
**APPENDIX F(6): STRATIFIED SUPPORT SAMPLES FOR EACH CLASS IN A BALANCED TEST SET OBTAINED USING SMOTE – ENN METHOD IN FIVE FOLDS FOR DIASTOLIC FUNCTION**

****

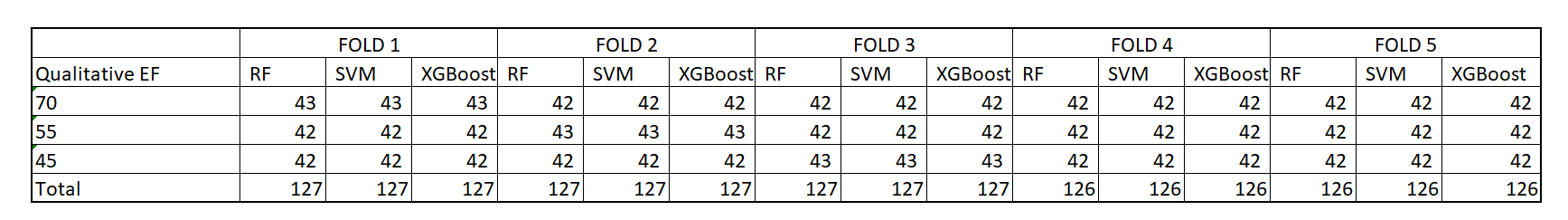
**APPENDIX F(7A): STRATIFIED SUPPORT SAMPLES FOR EACH CLASS IN A BALANCED TEST SET OBTAINED USING SMOTE – ENN METHOD IN FIVE FOLDS FOR EJECTION FRACTION (NUMERIC)**

****

**APPENDIX F(7B): STRATIFIED SUPPORT SAMPLES FOR EACH CLASS IN A BALANCED TEST SET OBTAINED USING SMOTE – ENN METHOD IN FIVE FOLDS FOR EJECTION FRACTION (RANGES)**

****

**APPENDIX F(7C): STRATIFIED SUPPORT SAMPLES FOR EACH CLASS IN A BALANCED TEST SET OBTAINED USING SMOTE – ENN METHOD IN FIVE FOLDS FOR EJECTION FRACTION (QUALITATIVE)**

****

**APPENDIX G: FLEISS KAPPA SCORE FOR INTER-RATER RELIABILITY USING R STUDIO 1.2.5033**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **#** | **PRIMARY OUTCOMES** | **NO OF SUBJECTS** | **NO OF RATERS** | **FLEISS'S KAPPA SCORE** |
| 1 | MITRAL REGURGITATION | 37 | 3 | 0.85 |
| 2 | MITRAL STENOSIS | 51 | 3 | 0.93 |
| 3 | AORTIC REGURGITATION | 30 | 3 | 0.81 |
| 4 | AORTIC STENOSIS | 25 | 3 | 0.83 |
| 5 | FLOW GRADIENT | 25 | 3 | 0.83 |
| 6 | DIASTOLIC FUNCTION | 25 | 3 | 0.83 |
| 7 | EJECTION FRACTION | 67 | 3 | 0.91 |

‘irr’ and ‘lpsolve’ packages have been used in R to compute the Fleiss kappa score