APPENDIX A

1.1 Granularity, Applicability, Prediction Techniques and, Evaluation Metrics

To summarize each paper selected in our final list, we present the granularity, applicability, prediction technique, and evaluation of the metrics found in each paper. We identified each paper in our final list using p[number]. The complete references can be accessed at https://github.com/igorwiese/esem14 (references.pdf).

The granularity represents which are the units of study in prediction models. For example, some papers built prediction models to predict defects on the file, while other on packages/modules. Table 1 presents the granularity of each study. Observing the applicability, we noticed that more than a half of the papers identified on our final list of papers built models to predict software fault. The most recent paper found on forward step was published on 2011. This result shows that new papers tend to explore different applicability. Instead of build models to predict failure, they focus on data gathered from issues to build models to predict tasks for bug triage, bug fixing time, re-opened issues and, severity of an issue.

Table 1. Granularity of studies

Granularity	Seed	DBLP	Backward	Forward
File	p1,p7,p10,p12, p15,p17,p18, p19,p20,p24, ,p26	p40, p41,p42,	p30,p32,p33, p34,p35,p36, p37, p39, p43	p44,p50,p51, p52,p53,p54,
Commit	p2			
Module	p3,p9,p10,p16, p22,p27, p28		p29,p34	
Binary	p5,p6,p27		p38	
Build	p8,p13			
Thread mail	p21			
Issue				p45,p46,p47, p48,p49,p51

Table 2 shows the applicability of each study. We used the applicability to highlight which are the prediction goal of each study. In this sense, we consider granularity on file level and the applicability related to fault prediction. We found that 60.41% of papers used the file and, 18.75% considered module as a granularity

Table 2. Applicability of studies

Applicability	Seed	DBLP	Backward	Forward
Fault	p1,p3,p5,p6,p7, p9,p10,p15,p16, p17,p20,p22,p24, p26,p27,p28	,p40,p41,p42	p29,p30,p32,p33,p34,p35,p37, p38 p39,p43	p44,p50,p51,p53,p54
Risky	p2,p18			
Build	p8,p13			
Vulnerability	p12,p19		p36	
Discussion	p21			
bug triage				p45,p49
bug fixing time				p46,p48
re-opened Issues				p47,p49
Effort				p51
Severity				p51
Churn of file				p52

Since we have different granularity and applicability's we also summarize which techniques were used to build prediction models. We found that two types of prediction models: Classification and Statistical. Table 3 presents the classification algorithms used by each paper. It is possible to observe that different amount of algorithms were used in the same study. The range varies from 1 to 4 algorithms used. Naive Bayes was the most frequently algorithm used (11 papers).

Table 3. Classification techniques

Model	Seed	DBLP	Backward	Forward
Naive Bayes	p1,p7,p13,p20, p21,p26	p40	p34,p36	p46,p49
J48	p1,p20		p36,p37	p44,p48
Decision tree	p3,p21,p26		p30	p44,p46, p52
Bayes Network	p12,p19		p36	p44,p46, p48
SVM	p20			p45,p49
Random Forest			p36	
RBF Bayes				p46,p48
Neural Networks				p52
KNN				p46

Considering the statistical models applied (Table 4), we found that seven different types of regression models were used. The logistic regression was selected by 43.63% of papers, this model was the most popular prediction model, considering statistical and classification models. We found papers that used both classification and statistical models to evaluate their approaches, for example, p[1], p[3], p[20], p[26], p[30], p[36], p[44] and, p[48].

Table 4. Statistical techniques

Model	Seed	DBLP	Backward	Forward
Logistic Regression	p1,p2,p3,p6,p8, p9,p10,p15,p16, p17,p18,p20,p22, ,p26		p32,p36,p38, p43	p44,p47, p48
Linear Regression	p3		p30	p48
Binomial Regression	p5, p15, p24,p28	p41,p42	p33	
Multiple linear regression	p27		p39	p48,p50
Poisson Regression	p15		p29	
Regression tree(m5)			p30	
Non-linear Decision tree			p35	

Table 5 shows the evaluation metrics applied by each study. We found that recall, precision and AUC (area under the curve) were the most popular metrics performed to evaluate classification algorithms. Each paper normally performed more than one evaluation metric.

Table 5. Evaluation metrics applied to Classification

Evaluation	Seed	DBLP	Backward	Forward
Recall	p2,p5,p7,p9,p10,p12,p13,p16,p19, p20,p21,p22			p44,p45,p48,p49
Precision	p2,p5,p7,p9,p10,p12,p13,p16,p19, p20,p21,p22			p44,p45,p48,p49,p54
F-measure	p16,p20			p44,p45,p46,p49
Cost-Sensitive	p1,p20			
AUC – ROC	p3,p7,p10,p16, p19,p26,p20	p40	p35,p37	p47,p48
Accuracy	p5,p20,p21			p49,p54
Cost-Curve	p7,p10,p34			
Lift Cumulative Chart	p15,p26			
Effort aware analysis	p26			
Misclassification Rate	p21			p54
Type I and Type II	p18			
Inspection Rate	p19		p36	
Probability of Detection (PD) and false alarm (PF)			p34,p36	

Table 6 presents the evaluation metrics applied to statistical models. The most popular metrics selected were R^2 adjusted, Mean Absolute Error (MAE) and correlation between the independent variables with dependent variable. Again, we observed that more than one metric were frequently selected to evaluate the prediction models.

Table 6. Evaluation metrics applied to Statistical models

Evaluation	Seed	DBLP	Backward	Forward
R^2 adjusted	p2,p8,p9,p16,p17,p18,p27		p32,p38, p39, p43	p48,p50
Beta coefficients	p6,p8	p41,p42	p29,p32	p50
Pearson Correlation	p9		p37	p52
Spearman Correlation	p9,p12,p15,p22, p28		p29,p30, p35,p37, p38, p39	p44,p47,p48,p51,p53
Deviance (D2)			p32, p43	
AIC Scores				
Odds Ratio	p17			p47
Mean Absolute Error			p30,p35, p37	p44,p50,p52
Mean square error			p30	
Root mean Square Error			p35,p37	p44
Discriminative Power (Welch test with bonferroni)			p36	
Mean FPA		p41,p42		
Mean root square				p46
Granger Causality Test				p51
Kendall Correlation				p52

Root mean square Deviation				p52
Mean Relative Error				p52

We observed a lack of papers reporting the use of feature selection algorithms or effect size analyses trying to explain the contribution of each individual predictor.

Finally, we mapped which projects were object of study to build prediction models. We identified 89 different projects. We found 21 commercial projects, like Windows Vista and Rational team concert. We mapped 69 open source projects most selected. Eclipse and Firefox were most frequently studied. The list of projects contains Eclipse (13), Firefox (7), Windows vista (7), Lucene (5), Rational team concert (4), CXF (3) and, Wicket (3).

Considering all 89 projects, 11 appeared on more than one-step of our systematic review. The projects are: Rational team concert (seed and forward), Firefox (seed, backward, forward), Netbeans (backward, forward), Eclipse (seed, backward, forward), Mylyn (seed, forward), Eclipse PDE Build and UI (seed, backward), Equinox (seed, dblp), Lucene (seed, dblp), Qpid (seed, forward), VLC (forward, dblp).