

Classificação de código Fonte Utilizando Características de Textura de Code Minimap

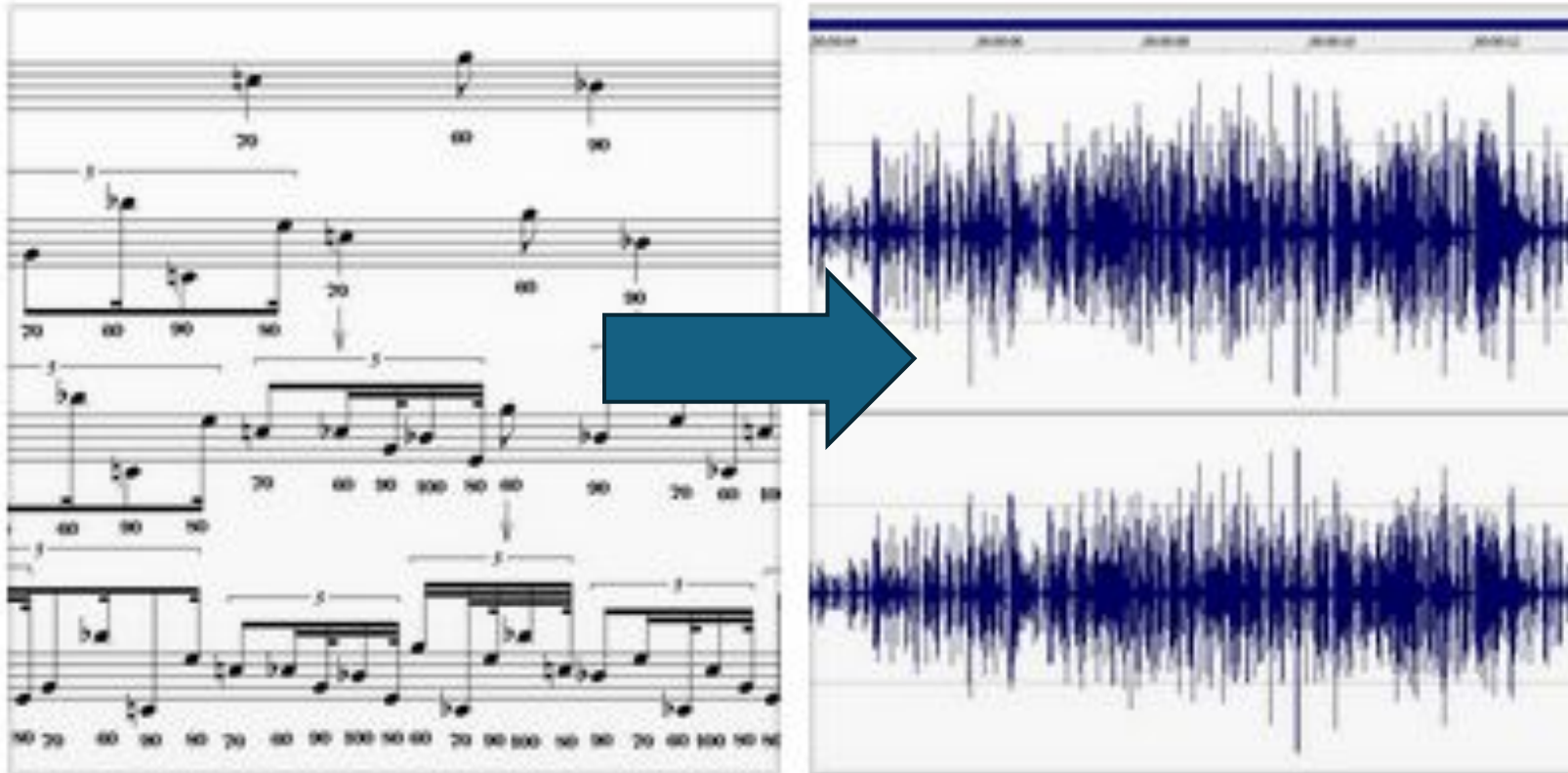
Munif Gebara Junior, (UEM)

Dr. Yandre Maldonato e Gomes da Costa (UEM)

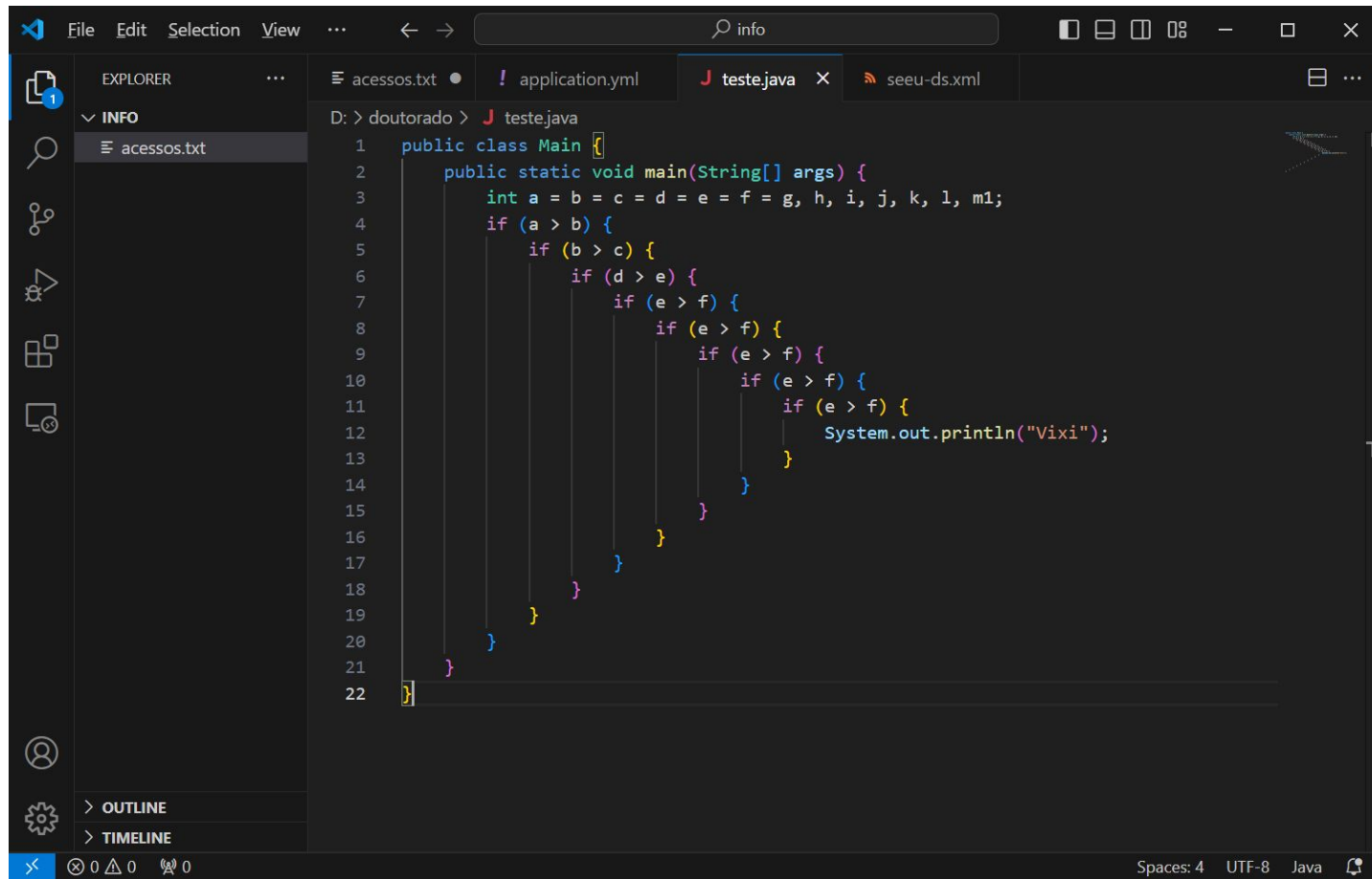
Dr. Igor Wiese, (UTFPR)

Referencial teórico

- COSTA, Y. M. G.; Oliveira, L.S.; Koerich, A.L.; GOUYON, F.; Martins, J.G. Music
- Genre Classification Using LBP Textural Features



MiniMap



The screenshot shows the Visual Studio Code editor interface. The Explorer sidebar on the left displays the file structure, including 'acessos.txt' and 'INFO'. The main editor window is open to the file 'teste.java', which contains the following Java code:

```
D: > doutorado > J teste.java
1 public class Main {
2     public static void main(String[] args) {
3         int a = b = c = d = e = f = g, h, i, j, k, l, m1;
4         if (a > b) {
5             if (b > c) {
6                 if (d > e) {
7                     if (e > f) {
8                         if (e > f) {
9                             if (e > f) {
10                                if (e > f) {
11                                    if (e > f) {
12                                        System.out.println("Vixi");
13                                    }
14                                }
15                            }
16                        }
17                    }
18                }
19            }
20        }
21    }
22 }
```

The MiniMap view is visible on the right side of the editor window, providing a visual overview of the code structure. The status bar at the bottom indicates 'Spaces: 4', 'UTF-8', and 'Java'.



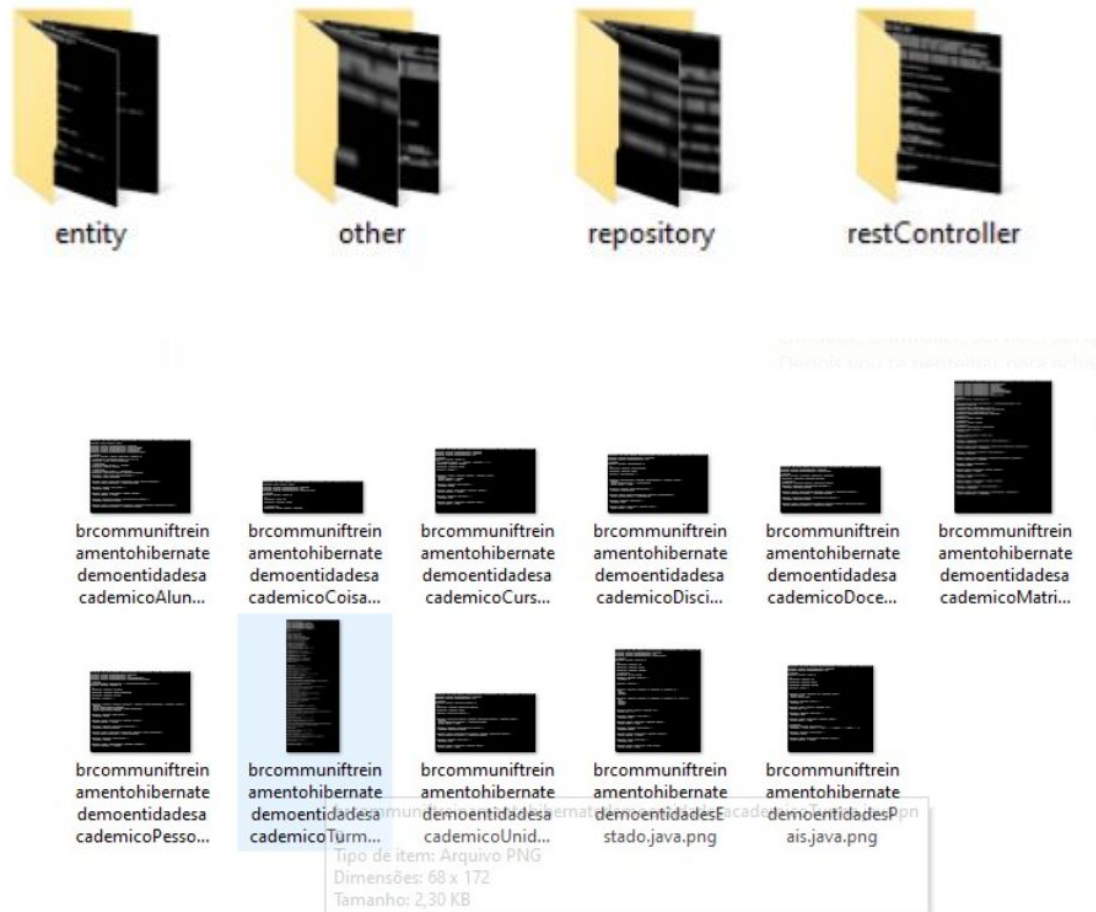
Criação do Dataset automatizada

```
with open(caminho_arquivo, 'r', encoding=encoding) as arquivo:
    for linha in arquivo:
        for x in range(len(linha.strip())):
            v = ord(linha[x])
            v = encrypt_chars(v)
            if v > 255:
                v = 255
            try:
                imagem.putpixel(xy: (x, y), value: (v, v, v))
            except IndexError:
                print(largura, altura, x, y, v)
        y += 1

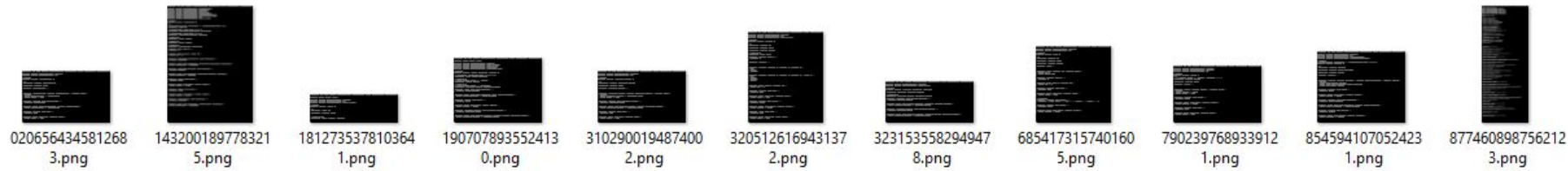
caminho_saida_com_tipo = caminho_saida + '/' + tipo + '/'
os.makedirs(caminho_saida_com_tipo, exist_ok=True)
arquivoImagem = create_16_digit_hash(caminho_arquivo) + '.png'
imagem.save(caminho_saida_com_tipo + arquivoImagem)
```

Criar um Dataset classificando Estereótipos

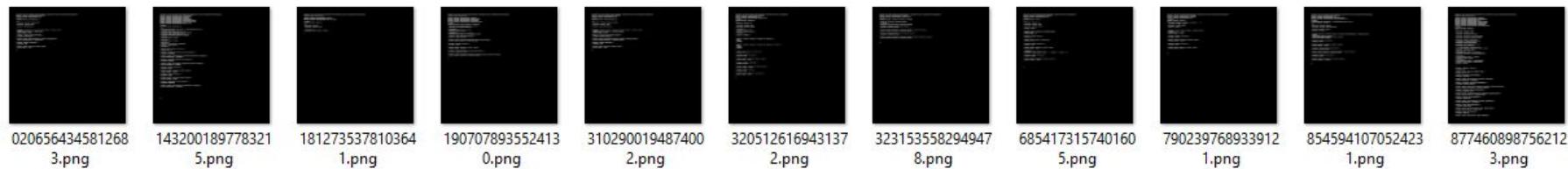
Fase Atual



Padronizar tamanho da saída



- 128px 128px
- Bordas 8px
- Área com dados 128px-8px-8px 112px 112px



Impedir engenharia reversa do código fonte

- Hash do nome dos arquivos
- Código ASCII transformado



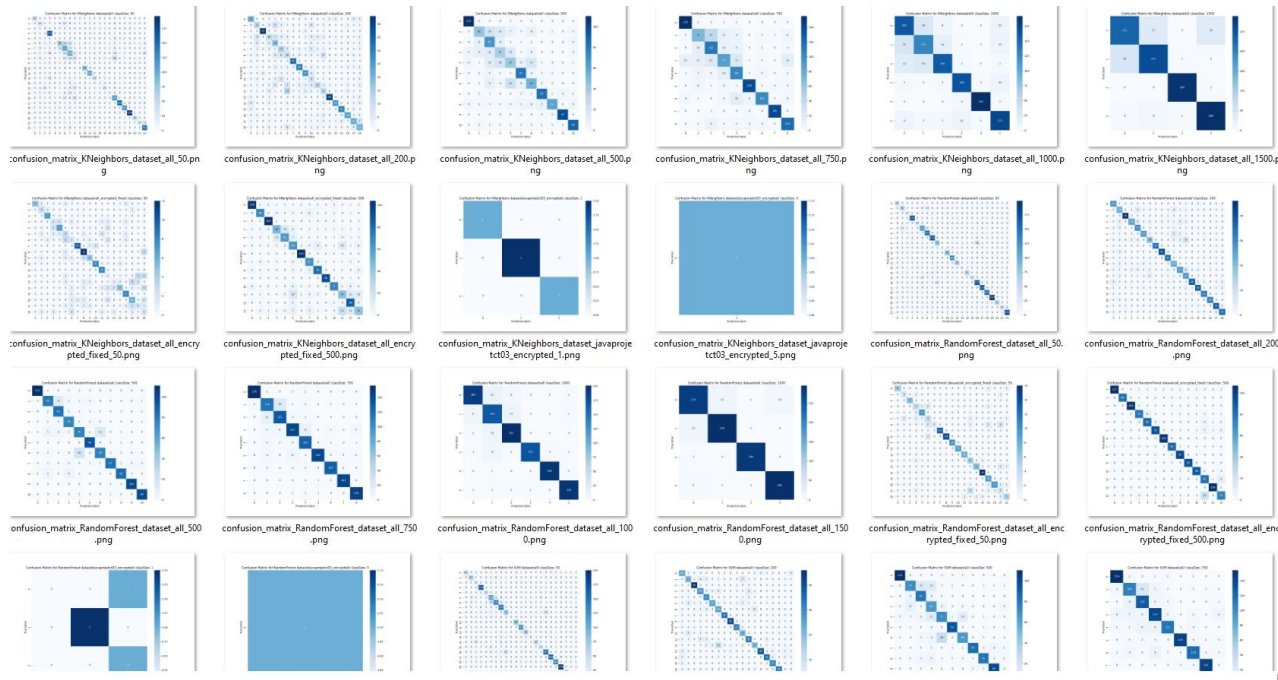
```
def encrypt_chars(asc_code): 1 usage new *  
    if asc_code < 33: # não-imprimíveis  
        return 32  
    if asc_code < 48: # símbolos  
        return asc_code  
    if asc_code < 58: # números  
        return 53  
    if asc_code < 65: # símbolos  
        return asc_code  
    if asc_code < 91: # maiúsculas  
        return 77  
    if asc_code < 97: # símbolos  
        return asc_code  
    if asc_code < 123: # minúsculas  
        return 109  
    if asc_code < 127: # símbolos  
        return asc_code  
    return 130
```


Base de Dados

14 Classes

500 amostras em cada classe]

<https://github.com/munifgebara/codeminimap>



0 css

1 html

2 javaconverter

3 javadto

4 javaentity

5 javaimplementation

6 javajasper

7 javajsp

8 js

9 json

10 sql

11 javaintegrationtest

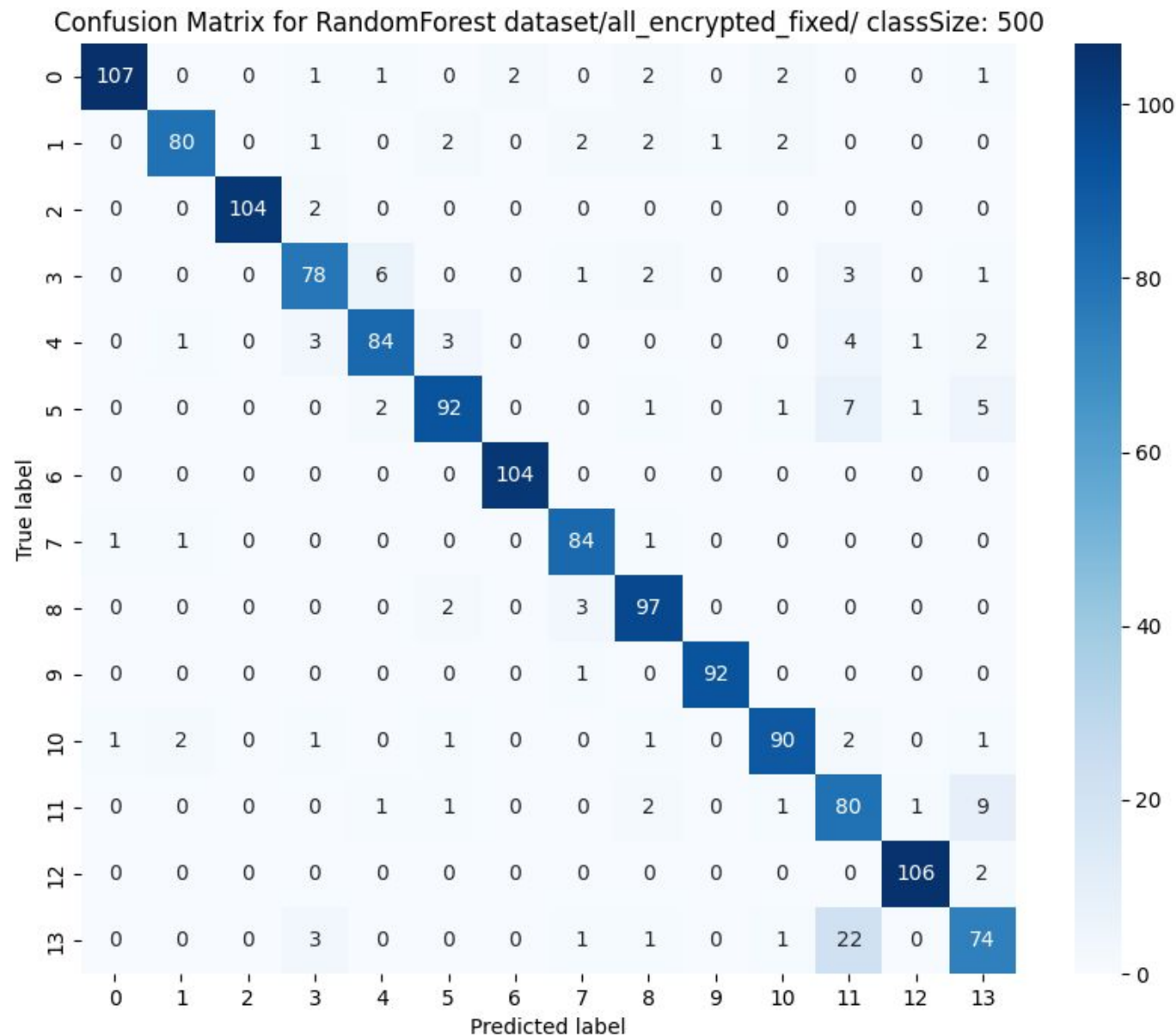
12 javajsf

13 javaunittest

Resultados

0 css
1 html
2 javaconverter
3 javadto
4 javaentity
5 javaimplementation
6 javajasper
7 javajsp
8 js
9 json
10 sql
11 javaintegrationtest
12 javajsf
13 javaunittest

accuracy 0.91
macro avg 0.91
weighted avg 0.91



Próximos passos

- Criar outros datasets:
 - Repositórios open source e privados
 - Outras linguagens
 - Outras classificações
 - Outros tipos de textos
 - Comparar com outros métodos
 - Responder: “Tais métodos são promissores para análise de código fonte ?
 - Testar deep learning
- Publicar dataset e artigo

Feature Detection in Software: A Source Code Mini-Map Approach

Decoding Software using Sourcecode mini-maps to identify programming Languages and stereotypes

1st Munif Gebara Junior

DIN

UEM - State University of Maringá UEM - State University of Maringá UTFPR - Federal University of Technology (Brazil)
Maringá, Brazil
pg55752@uem.br

2nd Yandre Costa

DIN

Maringá, Brazil
ymgcosta@uem.br

3rd Igor Wiese

??????

Paraná, Brazil
igor.wiese@gmail.com

Abstract—This document is a model and instructions for \LaTeX . This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. ***CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

Index Terms—component, formatting, style, styling, insert

I. INTRODUCTION

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
- Avoid combining SI and CGS units. such as current