ISO/IEC TR 17903:2024-05 (E)

Information technology - Artificial intelligence - Overview of machine learning computing devices

Contents

Page

Foreword iv

Introduction v

1 Scope 1

2 Normative references 1

3 Terms and definitions 1

4 Abbreviated terms 5

5 ML computing device concepts 5
5.1 Processing 5
5.2 Computing 5
5.3 Device 6
5.4 Infrastructure 6
5.5 Service 6
5.6 Performance 7
5.7 Computing device 7

6 ML computing device characteristics 7
6.1 Datatypes 7
6.1.1 General 7
6.1.2 Effectiveness and efficiency 8
6.2 ML operators 8
6.2.1 General 8
6.2.2 Effectiveness and efficiency 9
6.3 Memory access and addressing mechanisms 10
6.3.1 General 10
6.3.2 Effectiveness and efficiency 10
6.4 Scheduling 10
6.4.1 General 10
6.4.2 Effectiveness and efficiency 11
6.5 Topologies 12
6.5.1 General 12
6.5.2 Effectiveness and efficiency 12
6.6 Streams 13
6.6.1 General 13
6.6.2 Effectiveness and efficiency 13
6.7 Buffering mechanisms 13
6.7.1 General 13
6.7.2 Effectiveness and efficiency 14
6.8 Cache mechanisms 14
6.8.1 General 14
6.8.2 Effectiveness and efficiency 14
6.9 Data exchange mechanisms 15
6.9.1 General 15
6.9.2 Effectiveness and efficiency 15

7 Approaches and measures for performance optimization 16

7.1 Approaches 16
7.1.1 Overview 16
7.1.2 Computing resource-level 16
7.1.3 Enabling software-level 16
7.2 Measures 17

Annex A (informative) Relationships between ML computing device-related definitions 19

Bibliography 21