**a) File name of the graphical abstract:**

graphical\_abstract.pdf

**b) Caption for the graphical abstract:**

This paper presents a scalable hardware architecture for Spike-by-Spike neural network computation in embedded systems. The key contribution of this research is the design of a dot-product hardware unit based on approximation techniques. This approach leverages the intrinsic error resilience of neural networks to accelerate computation, reduce memory footprint and power dissipation while preserving inference accuracy.